lady pupils, ranging from thirteen or fourteen years of age t
,How old are you?' ,Over thirteen sir,' replied the child. ,Oh!
ng boy of about twelve or thirteen very subject to that restle
dy money (amounting to thirteen pounds and half-a-crown
he heels of his boots, for thirteen minutes by the college c
him, ,Or a fellow of about thirteen or fourteen stone might b

TOKEN

A Journal of English Linguistics

Volume 13/2021

anxieties and shames. At thirteen she could read and keep
have a taste that way. At thirteen years old, the Child of the
usual wages of a boy of thirteen thirteen; and likewise faile
nfortunate, who might be thirteen years old, and was lookin
en old Luke Withers won thirteen times running on the red
ng as I have, sir, and had thirteen children of her own, and
pened by a young girl of thirteen or fourteen. The undertak
ld. Twelve, or at the most thirteen might be near the mark.
re's a pair of Hessians in thirteen there's two pair of halves
r's little servant, who was thirteen years old and three feet h
hat she had begun to be thirteen years ago, and Sir Walter
hbourhood afforded, and thirteen springs shewn their blos
ngs soon. That will make thirteen with ourselves, so there v
her sense, she did not, at thirteen bid fair to equal her siste
Three months comprised thirteen weeks. Much might happ
parts of the work which thirteen years have made compa

ISSN 2719-4418

Token: A Journal of English Linguistics

Volume 13

JAN KOCHANOWSKI UNIVERSITY OF KIELCE

Token: A Journal of English Linguistics

Volume 13

Edited by

John G. Newman
Marina Dossena
Sylwester Łodej

Jan Kochanowski University Press

Kielce 2021

Cover design Jakub Patryk Łodej, Anna Domańska

Formatting Marzena Buksińska

Editorial Correspondence
Jan Kochanowski University
Institute of Linguistics and Literary Studies
ul. Uniwersytecka 17
25-406 Kielce
Poland
www.ujk.edu.pl/token

The *Capsula eburnea* in several Middle English witnesses

Isabel de la Cruz-Cabanillas* and Irene Diego-Rodríguez**

* *Universidad de Alcalá*
** *Universidad Antonio de Nebrija*

ABSTRACT

Medieval treatises containing predictions to recognise the signs of death were based on works written by Hippocrates or attributed to him. In the case of the *Capsula eburnea*, the original text was written in Greek and translated into Latin in the Middle Ages. The Latin translations circulated widely in different versions: a translation from Greek between the fifth and the seventh centuries and a later translation from Arabic in the late twelfth century. During the late Middle English period, they were translated into English, among other vernacular languages. The present article aims to compare and collate four fifteenth-century prognostic treatises in Middle English with their possible Latin exemplars. The analysis of the witnesses will shed light on the shadowy landscape of pseudo-Hippocratic prognostic texts in Middle English and will contribute to trace the Latin sources of these Middle English witnesses.

Keywords: *Capsula eburnea,* Middle English, Hunter 513, Additional 34111, Cambridge Dd VI 29.

1. Hippocratic prognostic treatises

Medieval physicians and monks working with patients were instructed in the recognition of the signs of death (Arrizabalaga 1999: 243; Paxton 1993: 631). Their predictions were mainly based on several works allegedly written by classic physicians like Hippocrates or Galen. These treatises containing signs foretelling death are included in medical compilations during the Middle Ages and transmitted along with other medical works.

DOI: 10.25951/4818

They were considered useful tools to let sick people know when death was near, because the physician would gain credit among their patients and professional prestige regardless of the result (Kuhne 1989a: 3). These reasons explain why a wide variety of Greek and Arabic sources dealing with the signs of death were firstly translated into Latin, and later, during the late medieval period, into different European vernaculars.

Most of these treatises were attributed to Hippocrates, even if they were written in the Middle Ages. The tradition of these pseudo-Hippocratic prognostic texts is obscure, since there is a series of related tracts known as *Letter of Ipocras*, *Capsula eburnea, Signa Vitae et Mortis, Treatise on Apostemes* and even *Analogium, Liber Praestantiae* and *Liber de Veritate or Liber Veritatis*. Sudhoff (1914: 81) illustrates the confusion surrounding the different denominations and includes even more titles to refer to these pseudo-Hippocratic prognostic treatises:

> But what is "Liber veritatis ypocratis tractatus unus?" Perhaps the "Capsula eburnea", perhaps the other prognostic treatise of the Apostemes: "Prognosticorum liber, qui dicitur liber secretorum" or the further pseudo-Hippocratic astrological treatise, "Libellus divinus de esse aegrorum secundum lunam"?[1]

In addition, at times, the information to determine the time of expected death is based on the features present on the patient´s face and this explains why in Latin it is also known as *Facies Hippocratica* (Kibre 1982: 177). The Latin manuscripts may even include other denominations, such as *Secreta Hippocratis* from the thirteenth century onwards (Sudhoff 1915/16: 82; Muschel 1932: 44; Kibre 1982: 178; Kuhne 1984/85: 32), *Analogum*, *Liber Praestantiae*, *Liber de Veritate* or *Liber Veritatis* (Kibre 1978: 194), *Prognostica*, *De Pustulis*, *Secreta*, *Signa Vitae et Mortis* (Kibre 1978: 194), *Prognostica Democriti* (Sudhoff 1915/16: 81; Sigerist 1921: 157). This terminological confusion, along with the fact that it is a short treatise inserted in bigger medical writings (Sigerist 1921: 157), makes it impossible to establish the real number of versions that have come down to us. For instance, Kibre (1978) includes one hundred and twenty-three extant copies in her list of Latin texts.

[1] Authors' translation of the original: „Was aber ist „Liber veritatis ypocratis tractatus unus"? Vielleicht die „Capsula eburne", vielleicht der andere prognostische Traktat aus den Apostemen: "Prognosticorum liber, qui dicitur liber secretorum" oder der fernere pseudohippokratische astrologische Traktat "Libellus divinus de esse aegrorum secundum lunam?".

The wide variety of denominations will give the reader an idea of how arduous the task of tracing the transmission of the text under study is. Our interest is in the *Capsula eburnea*, but very often it can be found in catalogues under one of the above-mentioned names. The localisation of the treatise is complicated not only because of the different titles used, but also because it is a very short tract (one thousand words approximately) and therefore, some manuscripts fail in the distinction of one work from another, as they appear one after another in such a way that there is no clear division between texts. These reasons make the tract invisible in catalogues at times, either not mentioning its presence or acknowledging it with a vague label like "Medicina Mortis et Vite; Ex Ipocrate | et Galieno, ut videtur |" (Glasgow University Library Catalogue, Hunter 323).

The Latin tradition of the text has been widely explored by Sudhoff (1915/16) and Kibre (1978), sources of all other pieces of research after them. Their contribution to the establishment of the Latin stemma is valuable, even if it is not without minor problems. For instance, Kibre (1978: 204) claims Escorial L.III.30 is a fifteenth-century copy, even if our examination of the manuscript shows that it is a seventeenth-century witness, which reads "MANVSCRIPTVS MEDICVS. ANNO M.DC.LXX.VI. Para la R' Cassa dl S. Lorenzo; Scorial" on the first page.[2] Dealing with such a vast number of sources may have hindered Kibre from consulting all the manuscripts and she must have relied on the information provided by the catalogues available to her at the time, which did not always offer an accurate date.

Several scholars include the vernacular copies produced in the different European languages. Thus, Beaujouan (1972: 187) and Alvar Ezquerra (2001: 46) mention the translation into Spanish. Similarly, whereas Meyer (1903) analyses one of the French versions, Benati (2013) and Di Clemente (2011) deal with the Middle Low German copies of the treatise and, in turn, Kibre (1945: 391 and 1978: 195) provides information on French, German, Italian and English translations. Finally, Voigts and Kurtz (2000) refer to English versions exclusively. This source is the only one which uses the denomination *Tokens of Ipocras* for the treatise. Our aim is to draw special attention to this specific Middle English treatise, which has passed unnoticed in the eyes of academia, and whose existing copies are grouped under the common designation of *Signa Vitae et Mortis, Tokens of Ipocras* or *Capsula eburnea*.

[2] Our special thanks to José Luis del Valle Merino from the Royal Library of the Monastery of El Escorial, who made the photographs of the manuscript available to us.

2. The transmission of the *Capsula eburnea*

Before concentrating on the Middle English manuscript copies analysed in this article, a summary of the transmission of the text is needed to clarify the origin of the English version. Sudhoff (1915/16) and other scholars following him (Muschel 1932; Kibre 1978; Benati 2013) trace back the origin of this treatise to the fourth or fifth century, when it was composed in Greek in the Eastern Mediterranean area. Later, between the fifth and the seventh centuries, it reached the South of Italy in a Latin translation. In the seventh century it was also translated from Greek into Arabic. This translation reached the Iberian Peninsula during the early Middle Ages and it was retranslated into Latin by Gerard of Cremona shortly after 1170 (Sudhoff 1915/16: 111). Gerard of Cremona used the denomination *Capsula eburnea* for his Latin translation of the Arabic version of the text and, for this reason, this title became standard only after the second half of the twelfth century. Both Latin translations spread around Europe, and "from the thirteenth century onwards, the *Capsula eburnea* is also witnessed in Old French, Middle English and Middle Dutch" (Benati 2013: 6). Apart from these two main recensions, Sudhoff also identified a third version from Greek and a fourth one that relied heavily on the Arabic one (1915/16). Sudhoff (1915/16) and Kibre (1978) identified manuscripts belonging to both traditions: the early anonymous Latin version from Greek and Cremona´s Latin version from Arabic.[3]

All these copies, which probably derive from the authentic *Liber Pronosticorum* (Kibre 1978: 194), show differences regarding their title, verbal content or as far as the attribution of the prognostic treatise to Hippocrates is concerned. In some manuscripts, the text appears under the names of other learned physicians: Democritus, Soranus or Galen (Sudhoff 1915/16: 86; Kibre 1978: 194). The structure of the *Capsula eburnea* also varies significantly in the tradition, although the general nature of their content is very similar: "They all deal with cutaneous eruptions as prognostic signs" (Benati 2013: 6), which

[3] Another revealing lead to be followed is the Hebrew transmission of the text through Arabic. In this respect, Muschel concludes that "the author of the Hebrew Capsula eburnea may have used an Arabic and not a Latin model" (1932: 59). [Authors' translation of the original: "der Verfasser der hebräischen Capsula eburnea wahrscheinlich eine arabische und nicht eine lateinische Vorlage benutzt hatte"]. Likewise, Kuhne (1989a, 1989b and 1990) has shown that the Arabic tradition was not made up of a single exemplar, but probably there was more than one source in the Arabic textual tradition. In fact, she claims that Sudhoff´s third recension is not from Greek as he contends, but from Arabic as well (1984/85: 37, 1986: 254 and 1989a: 5). Kuhne also concludes that the fourth recension is a summary of the second one but inspired by the Latin translation, not the original Arabic text (1984/85: 37).

means physicians would be able to identify these signs in their patients and, thus, foretell their death (Arrizabalaga 1999: 245). Some sources consist of "a title and a list of prognostic remarks" (Benati 2013: 7), whereas in other manuscripts an anecdotic introduction is found. In these cases, as explained by Kibre (1978: 194-195), the treatise includes

> A brief account in which Hippocrates is purported, when he was nearing death, to have ordered his retainers to place at his head in the tomb with him a small ivory box (*Capsula eburnea*) into which he had placed on an ´*Epistle*´ or receptacle containing the secrets of the medical art, and particularly those relating to the signs of life and death. At a later time, Caesar is said to have come upon the tomb and to have ordered that it be opened secretly. He thus found the receptacle resting under Hippocrates' head and requested that it be given to his own physicians. Henceforth, from the contents of this receptacle, the account concludes, physicians were able to learn and recognize the signs of life and death.

3. The Middle English tradition of the *Capsula eburnea*

The Middle English translations of the text have never been ascribed to any specific tradition. According to Robbins (1970: 287), a detailed inspection and continued search among Middle English medical manuscripts would no doubt uncover further texts. It is our aim to trace the original versions from which the English translations derive as well as to continue searching for more unveiled English texts.

3.1 Selection of the texts under study

Thus far no complete list of manuscripts containing the Middle English *Capsula eburnea* has been elaborated. Subsequently, localising the English manuscript copies of the *Capsula eburnea* is troublesome. By consulting all the different catalogues and critical works available to us (Young – Aitken 1908; Kibre 1977, 1978; Keiser 1998; Voigts – Kurtz 2000; online catalogue of the *Sloane Manuscripts* in the British Library), we have been able to establish a rough distinction between the different pseudo-Hippocratic treatises in prognostic matters. The information provided by Kibre (1978: 195) includes the following Middle English witnesses of the *Capsula eburnea*: BL Additional 34111, BL Sloane 405, BL Sloane 706 and BL Sloane 715. The first of the manuscripts containing two versions of the tract will be the subject of study here, along with two others. We have not had the chance to examine the Sloane 405 yet, but we have anaylsed the contents of Sloane 715 and Sloane 706. The inspection of Sloane 715 reveals that it is an alchemical text of only seven folios while the rest are blank. As for Sloane 706, the catalogue of the British Library provides the following information: folio 95 "Hippocrates: Le liures que io Ypocras enoiai a Cesar: 14th-15th cent.: Engl.". Our examination of the text beginning with "This book ypocras sente vn to kynge Cesar" allows for the conclusion that it does not include the *Capsula eburnea*. What is found in this manuscript is another medical text known as the *Letter of Ipocras*, which is often confused with the *Capsula eburnea*.

In fact, Tavormina (2007: 633) mentions Sloane 706 in his study of the *Letter of Ipocras*. The similarities in the contents of these two pieces lie in the fact that both writings are in a letter format written under Hippocrates´s name and addressed to Caesar. As accounted by Hunt (1990: 100), the *Letter of Ipocras* was assembled in the Middle Ages. It usually begins with an introduction followed by the treatment of urines and concluding with a collection of medical recipes. The fact that it is entitled *Letter* makes it likely to be confused with the *Capsula eburnea*, since according to the introduction to the treatise in many extant copies, it is an epistle written by Hippocrates when he was about to die and who ordered to have it placed under his head in his tomb. Caesar found this letter in an ivory casket and sent it to his own physician, who is named Panodosius, Poamonodonosis, Proamodosio, Monodorus, Misdos or other alternating names depending on the manuscript. The contents of both are clearly differentiated, since the *Capsula eburnea* contains signs of death based on skin wounds or apostemes, which explains why this tract is sometimes referred to as *Treatise on Apostemes*.

We have also made use of Voigts – Kurtz (2000) for the selection of the texts. Several searches were launched and no results were obtained under *Capsula eburnea*, *Letter of Ipocras*, *Tokens of Ipocras*, *Signs of life and death*, *Signis mortis* and several others. The search under Hippocrates as an author retrieved forty-six items. Several of them are not related to the *Capsula eburnea*, but even those which are connected to it are not easily recognisable, since, for instance, BL Additional 34111 is referred to as *Secreta Ipocratis*. With the *incipit* "Whoso hath dolor" and *Tokens of Ipocras* as title, Voigts – Kurtz list three other items: Magdalen College Oxford 221, BL Sloane 405 and

Huntington, HM 64. Finally, with no title but also under the incipit "who so hath dolor", CUL Dd VI 29 is found. Likewise, our research on another pseudo-Hippocratic treatise, *Þe Booke of Ypocras* (De la Cruz – Diego 2018), led to the discovery of the *Treatise on the Signs of Death* in Glasgow University Library, Hunter 513, which contains the *Capsula eburnea* (Young – Aitken 1908: 421).

Additionally, other sources were consulted, such as the British Library Sloane Collection, where several manuscripts seem to be related to this specific piece. Apart from the above-mentioned Sloane 405, Sloane 706, and Sloane 715, under (*De*) *Signis Mortis*, other manuscripts are found; namely, Sloane 282, Sloane 284, Sloane 2320, Sloane 3531 and Sloane 3550. The fact that the title is in Latin is misleading, since it can correspond to texts written in Latin or in English. In fact, in the case of Sloane 3550 the British Library catalogue states the main language is English, but our examination of this specific piece shows it is in Latin. Likewise, even if the catalogue claims that Sloane 284, Sloane 2320 and Sloane 3531 include *De Signis Mortis* in fourteenth- fifteenth-century English versions, our examination reveals the texts are in Latin, as well. Finally, Sloane 282 also contains the *Capsula eburnea* in Latin.

Therefore, a complete list of English versions of the *Capsula eburnea* is still wanting. From a hypothetical collection containing seven versions (British Library Sloane 405, Oxford Magdalen College 221, Huntington HM 64, Glasgow University Library Hunter 513, BL Additional 34111 – including two versions of the treatise – and Cambridge University Library Cambridge Dd VI 29), for the present article we have concentrated on the last four fifteenth-century witnesses, some of which are unexplored thus-far:[4] GUL Hunter 513 (ff. 107r-109v), BL Additional 34111, which contains two versions of the treatise – version 1 (ff. 231r-233v) and version 2 (ff. 235v-238v) – and CUL Cambridge Dd VI 29 (ff. 30r-32r).

3.2 Analysis of the structure of the text

The texts have been transcribed and compared with the two recensions in Sudhoff (1915/16): (1) the anonymous Greek-Latin version and (2) the Arabic-Latin version by Gerard of Cremona. As Sudhoff´s texts cover up to twenty-one signs in the first recension and twenty-four in the second

[4] Di Clemente has kindly sent us her work on the versions of Additional 34111 yet to be published.

one, we have also supplemented the collation with some extracts from Arabic sources in Kuhne (1990: 56), which completed the twenty-four signs present in Sudhoff´s second recension and enlarged it up to thirty tokens. However, the comparison made it clear that the prognostications from signs twenty-five to thirty had nothing to do with the ones present in the Middle English texts. Finally, we have consulted Sigerist's transcription of Glasgow University Library Hunter 96. This witness, despite the similarities it shares with Sudhoff's first recension, cannot be considered a direct copy of it. Hunter 96 seems to be one of the earliest Latin translations that has come down to us, since the Glasgow University Library catalogue dates it to the eighth and ninth centuries and Kibre to the ninth and tenth centuries (1978: 196).

In the Middle English *Capsula eburnea* tradition, several parts in the witnesses are clearly distinguished: First of all, the beginning of the text. In Hunter 513 the text begins with the usual introduction: "Here begynnethe þe tokenys þat ypocras þe leche wrote to knowe the seke yf he myghte be hole thorughe medycyne", who ordered this document to be placed in his tomb.[5] Likewise, in the two copies extant in Additional 34111, it is attributed to Hippocrates who had it buried in his tomb. The versions in Hunter 513 and Additional 34111 are entitled *Secreta ypocratis,* a name that, according to Kuhne (1987/88: 432), corresponds to Sudhoff´s third recension, while the *Capsula eburnea* should be used to refer to Sudhoff's second recension. Thus, a similar introduction to the *Capsula eburnea* in Hunter 513 is found in the first version of Additional 34111 (fol. 231r):

> Here begynneþ þe priuetes of þe gode man. and. a wyse þat was yclepid ypocras þe whiche man whan he drew to deþe yclosed were þes priuetes in a case of euore and leyde þis case in his sepulcre wiþ him þat þes same priuetes ne shulde beo descouered among no man

and in Additional 34111, version 2 the beginning reads (fol. 235v):

> Now here bigynneþ ypocras his priuetes in a noþer maner þe whiche priuetes were ydo in a case of euore and leyde vnder his heued in his toumbe.

Likewise, Cambridge Dd VI 29 refers to the tokens written by Hippocrates, but no ubication of the document is provided (fol. 30r):

[5] For clarity sake, in the transcriptions all abbreviations have been silently expanded.

> Her begyns þe takyns. þat ypocras þe leche wrot. for to knaw þe seke. ȝif he miȝth be hool thorgh medicyn or noon.

Sometime later Caesar[6] found it in a *case of euore* (Additional 34111) or a *scrippe* (Hunter 513) under his head and had it taken to his own physician, named Amadas in the Hunter 513 copy. According to the *Middle English Dictionary*, the term *scrippe* was adopted from Old French *escharpe*, *escherpe*, *eskerpe*, *eschreppe*, *escreppe*, *escrip*(*p*)*e* and it translates as "bag or satchel". Both the *Middle English Dictionary* and the *Oxford English Dictionary* remark that it was used especially for the bag carried by pilgrims for alms, but it is not exactly a *box* or *casket of ivory*, as in the Additional 34111 versions. After this, the signs of death start with the sentence "Here begynnethe the tokens" in Hunter 513, which includes twenty-seven prognostic texts. The two versions in Additional 34111 also claimed the text to have been found by *Cesare the Emperoure* and, in the second version, even states this Caesar is *Julius Cesar*. This text was sent to other friends in the first version and no person is mentioned in the second one. The number of predictions differs from those in Hunter 513, being twenty-five predictions in the first version of Additional 34111 and twenty-six in the second version, which are followed by several recipes. In turn, the Cambridge Dd VI 29 lacks this part and, subsequently there is no mention of any addressee and some predictions are missing.

The number of tokens correlates with the length of the treatises: Hunter 513, being the longest, has 1,285 words, the second version in Additional 34111 has 1,232 words and Additional 34111 first version shows 1,143 words. Finally, the shortest tract is Cambridge Dd VI 29, which contains 824 words and follows the text in Hunter 513 very closely. As will be seen, it is not a *literatim* copy, as the wording is not alike. The Cambridge Dd VI 29 scribe may have had access to different exemplars, since part of the information provided in this text is not present in Hunter 513, but both manuscripts might have shared the same exemplars at some point.

3.3 Analysis of the contents of the tokens

Before proceeding to the examination of the tokens, it is worth highlighting other aspects present in the text. Firstly, the structure of the predictions

[6] "Seser þe Emperowre" in the Hunter 513 text and in other versions has usually been identified with Julius Caesar, although Kuhne (1989a: 7, note 15) notes that the Arabic version is *Qaysar*, a word that could mean any Roman Caesar.

follows a recurrent pattern that is found in all the versions, as described by Kuhne (1989a: 9-10), when referring to the Arabic tradition:

1. The first part of a conditional or temporal sentence beginning with *if* or *whan*

   a. Apart from the first sign in Hunter 513 and Cambridge Dd VI 29 and three of them in Additional 34111 version 2 that begin with *whoso*, the remaining tokens include the *if* clause, whereas all Additional 34111 version 2 signs start with *whan*, with one exception beginning "Now þat it is vpon þe veyn". This is relevant when having a look at Sudhoff´s versions, since the second one starts its sentences with *quando* (*when*), whereas the first one employs *si* (*if*) instead.

   b. The part of the body that is affected is mentioned.

   c. A skin affection where the word *pustule* is pervasive and its description according to its details regarding size, colour, whether it is painful, etc. Apart from these, Hunter 513 offers other symptoms not related to the skin.

2. The second part of the conditional sentence. This clause includes details on the time of death, usually specified in the number of days that will go by before the death takes place.

3. The confirmatory sign. Here physiological symptoms like thirst, hunger, transpiration, elimination of urine or tools, yawn, sneeze, vomit or more skin affections can be found.

Secondly, in terms of the layout, only Hunter 513 presents their signs numbered in the outer margin from sign number six onwards, whereas the other Middle English versions mark the beginning of the tokens with a paragraph mark. There is some mismatch in the sequence of the signs, since the four Middle English witnesses contain a different number of tokens. The first sign clearly coincides in all the manuscripts under study with some slight variation, as can be seen in Table 1. Here the patient has some sort of swelling, tumour or aposteme in the face and picks his nose constantly. In all copies but Hunter 513 and Cambridge Dd VI 29, he also rests his left hand upon his breast. However, these two manuscripts coincide with Hunter 96 in adding the headache, while the other versions start directly with symptoms in the face.

Table 1. Comparison of sign number one in the ME witnesses of *Capsula eburnea*

Hunter 513	Add. 34111 Version 1	Add. 34111 Version 2	Cambridge Dd VI 29
Here begynnethe the tokens fyrste of þe hede who so haþe doloure or ache in his hede or swellynge in his vesage with owten redde and with the lyft and allway pykud his nose thrylles þe xxiiijti day he schall dye	¶ Whan in þe face of þe Sekeman ariseþ a posteme and nys noȝt y.found no touche and þe left honde yleyd vpon þe brest he shalle die at 13 dai and nameliche whan in þe bygynnyng of his sekenes he gropeþ hys nose þrilles	Now ȝif a man haue ache or swellyng in þe face wiþ outen cogh and legeþ his left honde vpon his brest and makeþ hym wonder bysy to pyke and scratteþ þe nose þrilles shal dye wiþ in a short tyme 13	¶ ffirst for þe hede ake.or swellinge in þe face wiþ out rode. and wiþ þe lift honde always piketh his nose thriles. in xiiij. day he he schal dye.

In turn, Table 2 illustrates the sign in the Latin translation.

Table 2. Comparison of sign number one in the Latin witnesses of *Capsula eburnea*

Sudhoff Version 1	Sudhoff Version 2	Hunter 96
(I) Si habuerit dolorem vel tumorem in facie sine tusse <et sine ullo dolore> et sinistra manus vel pectus seu nares assidue scalpserit in XXII die morietur.	(I) Quando in facie infirmi fuerit apostema, cui non inuenitur tactus, et fuerit manus eius sinitra posita super pectus suum, tunc scias quod morietur usque ad 23. dies, et precipue quando in principio sue agritudinis palpat nares.	(I) In caput dolorem habentis siccum tumores in faciem habuerit sine dolorem et sinistra manus pectus et naris sibi adsidue scapet ad XXXII dies moritur.

From here the focus will be on the contents of the tokens, mainly the parts of the body mentioned and symptoms that can predict the death of the patient. Whereas signs number two and three are quite similar in all the versions, in sign number four Hunter 513 and Cambridge Dd VI 29 show no contents related to both versions in Additional 34111. As can be seen in Table 3, version 1 of Additional 34111 follows Sudhoff´s recension 2, while version 2 is comparable to Sudhoff´s first recension. In the Latin translations, the position of the pustule is under the tongue in recension 1 ("sub lingua illi papula"). Thus, in the second version in Additional 34111 "a whelk vnder þe tong" can be read, whereas in the second Latin recension "super linguam pustula" is rendered as "vpon þe tong a kirnel" in the first version of Additional 34111.

Table 3. Comparison of sign number four in four witnesses of *Capsula eburnea*

Add. 34111 Version 1	Add. 34111 Version 2	Sudhoff Version 1	Sudhoff Version 2
¶ Whan þat it is vpon þe tong a kirnel as a tike þow shalle wyte he shalle dye þe same day and þis is þe tokenyng of the sekenes ate þe bygynnyng desireþ hote metes in here kynd	¶ ȝif þat it be a whelk vnder þe tong and desireþ water and þan and aecke a feuer in þis sekenes and ȝif swellyng be in þe grete to grete or smale in þe seuen day shalle dye	(4) Item qui una in causa fuerit, si sub lingua illi papula apparuerit sicut lenticula quasi modica sive lavacra aut vaporem desideraverit [et intus passionis febricitantia fuerint – et si in digitis pedum pollicis tumor niger vel modice fuerit, in VII die morietur][7]	(4) Quando fuerit super linguam pustula, sicut musca canina aut sicut granum pentadactili, tunc scias quod patiens eam morietur in die et huius est signum, quod desiderat in principio res calidas in suis naturis.

Signs five and six in the Latin versions and in Additional 34111 deal with pustules in the feet. This part is missing in Hunter 513 and Cambridge Dd VI 29, inasmuch as the information after the first three signs corresponds to number six in Sudhoff´s first recension:

(1a) Item in febre acuta si in stomacho seu in dextro pede pustellam habuerit in planta, non altam sed aequalem, deterrimum humorem tenentem, et nullum desiderium habuerit, in XXII die morietur (Sudhoff 1915/16)

[7] Hunter 96 version is similar at the outset but differs notably in the second half of the sign: "(IV) Hubula cui in causa fuerit sub lingua papula apparuerit sicut tisticulis porciunus et labagra siue uapura uenerit inicium passionis ipsius morietur".

(1b) /¶/ Aso yf the seke be in the feuer ageus and haþe an euyll stomake and in the ryghte foote or in þe lefte fote wax A wenne or in the sole of the fote so þat it be not to grete but evyn lyche and as colour as ynde and A party swellynge and no desyringe to mete þe xxijti day he schall dye /¶/ (Hunter 513)

(1c) ¶ And þe seke be in a feuere agu. and hath euil stomak. oþer in þe riȝth foot. or in þe lifth. or in þe sole of þe foot wex a wen. but þat hit be not gret. but euinlich. and hath colour as ynde. and aparty whellith. and no talant hath to meth. in þe. xxij. day he schal dye. (Cambridge Dd VI 29)

From here, following the information provided by each version is not an easy task, since it appears in different order and, subsequently, the correspondence between signs is not always linear. Kuhne (1989a: 6) points out the equivalences between Sudhoff's recensions. Thus, according to her, the order follows the pattern shown in Table 4:[8]

Table 4. Distribution of signs in Sudhoff's first and second recensions

Signs in Sudhoff version 1	7	8	9	10	11	12	13	14	20	16	17	18
Signs in Sudhoff version 2	8	9	10	11	12	13	14	15	17	18	20	22

Furthermore, the information presented in Latin can be found in the sign above or below in Middle English. Some of the texts are numbered by the scribe, while in others there is no numbering. What is easily observed is the fact that Hunter 513 and Cambridge Dd VI 29 are comparable to Sudhoff's first recension and sometimes agree with Additional 34111 second version, whereas the first version in Additional 34111 relies on Sudhoff's second recension on a regular basis. A case where Hunter 513, Cambridge Dd VI 29 and Additional 34111 version 2 seem to have been following Sudhoff's first recension or the Hunter 96 tradition can be seen in Table 5.

[8] Our own examination of the texts makes us disagree on the final equivalences of her sequence, whereby in Sudhoff's version 1 signs 20 and 18 correspond respectively to signs 17 and 22 in Sudhoff's version 2.

Table 5. Comparison of Hunter 513, Add. 34111 version 2, Camb. Dd VI 29 and Sudhoff's version 1 and Hunter 96

Hunter 513	Add. 34111 version 2	Camb. Dd VI 29	Sudhoff version 1	Hunter 96
(10) /¶/ Also yf ~~the too~~ A man be seke of the splene / and blede at þe nostrellis as come the xiiij day he schall dye	¶ ȝif þat þe splene haue grete ache and ariseþ in þe left honde white whelkes and comeþ out þek blode ate þe nose shalle dye wiþ in short while	¶ And a man be seke in þe splene. and blede at þe nose thirles [as] come. þe .iiij. day he schal dye.	(11) Item si splen doluerit et papulae albae in sinitra manu ei natae fuerint inpares, et si per narem sanguis quasi spumosus cucurreit, in XII die morietur.	(XI) Spleneticum si papule multe albe sinistra manum apparuerint et per nares sanguis espumosis exierit in XII die moritur

Whereas the Latin texts and the second version in Additional 34111 mention a white wen on the left hand, this symptom is missing in Hunter 513 and Cambridge Dd VI 29. In addition, some features are present in only these two texts in Middle English. As an example of the similarity, a comparison to *beans of Egypt* is mentioned in these two versions as well as in Latin:

(2a) Saniem ex quacunque parte excreantibus si macule nate fuerint, sicut solent per omne corpus in modum fabe ægyptie Li. die morietur (Sudhoff 1915/16, version 1)

(2b) Qui sanguinem uomen si maculem per omne corpus exierit in modum fabe egicie LII d<ie> mor<itur> (Hunter 96)

(2c) /¶/ And yf the seke Caste blood /and there waxe blacke spottes þorughe owte his. body and þe membris be swollen and ryse bladderis like benys of Egipte þat day he dyethe for sothe (Hunter 513)

(2d) And ȝif þe seke caste blod and blak spotteȝ. shewiþ thurghtout þe body. and men bris be neth. and rise bledders as it ware benes of egypte. þe ilke day he schal dye (Cambridge Dd VI 29)

In turn, Additional 34111 version 1 can be practically considered a total rendition of Sudhoff´s second recension. This fact can be seen in sign twelve, where the symptoms are nose bleeding, a white pustule on the right hand and the rejection of food. The second version in Additional 34111 mentions blood spitting, but none of the other features.

Table 6. Comparison of sign twelve in Additional 34111 and Sudhoff's second version

Add. 34111 Version 1	Add. 34111 Version 2	Sudhoff Version 2
¶ Whan þat þe blode renneþ from þe nose þrilles and draweþ to whitnesse or to rednes and sheweþ in þe ryght honde a lytel white kyrnel he shalle deye þe þird day and in þe bigynnyng of þe sekenes he coueteþ metes in alle maner	¶ Who so speweþ blode and ariseþ whelkes white ouer alle þe body as grete as a bene shalle dye þe same day	(12) Quando fluit sanguis a naribus trahens ad subalbedinem uel rufedinem et apparet in manu dextra pustula alba non dolens, scias quod morietur die tertio sue egritudinis, signum est quod omnino non desiderat cibum.

Contrariwise, the information in Sudhoff´s first recension neither agrees with this one nor with version 2 in Additional 34111, but with the symptoms in Hunter 513 and Cambridge Dd VI 29.

Table 7. Comparison of sign twelve in Hunter 513, Cambridge Dd VI 29 and Sudhoff's first version

Hunter 513	Camb. Dd VI 29	Sudhoff Version 1
(12) /¶ Also yf he haue Euyll in the bladder and þe flessche in þe lefte syde swelle and he may not slepe with in xv• dayes he schall dye	¶ And ȝif he haue euyl in þe bleddur. and þe flesch of þe lift sydy swelle. and he may not slepe. þe .xv. day he schal dye.	(12) Nescie dolor cum fuerit, si in sinistra parte rubores spissi fuerint sine dolore et olera desiderauerit cottidie xxv die morietur.

As shown above, Hunter 513 and Cambridge Dd VI 29 share several pieces not present in the other texts. Thus, in sign thirteen specific symptoms in male genital parts may be foretelling death. A sign in Sudhoff's version 2 mentions the same part of male's body in its sign eighteen, but the rest of the symptoms do not coincide at all:

(3a) /¶ Also yf þe seke haue grete maledye / in þe lendis and fallythe into þe yerde / aftyr swell vp into þe wombe / and comythe to the herte þe v. day he schall dye (Hunter 513)

(3b) ¶ And ʒif þe seke haue gret malady in his lendes. and falliþ in to þe ʒerde of man. and aftur swelliþ vp in to þe wombe. and comiþ riʒth to þe hert. þe .xv. day he schal dye. (Cambridge Dd VI 29)

(3c) Et accidit dolor quibus dam in preputio, id est in cute cooperiente uirgam. Cum ergo dolor accidit alicui, deinde apparet in cubito pustula fusci coloris. Cum ergo dolor accidit alicui, deinde apparet in cubito pustula fusci coloris. Scias quod morietur in .ix. die sue egritudinis ante solis ortum, et signum est, quod desiderat in principio sue egritudinis bibere vinum. (Sudhoff 1915/16, version 2)

The affinity between both Middle English manuscripts is even more noticeable in the final part, where the information presented is completely absent in the other texts. This divergence from the two Latin traditions (Sudhoff's version 1 and 2) and the English translations from it (Additional 34111) could be due to the influence of the third Latin recension that circulated at the end of the fifteenth century, according to Sudhoff (1915/16) and Kuhne (1984/5). Thus, it is unknown whether the following passages could be inspired by the third Latin tradition:

(4a) /¶ Also yf ther waxe mechill spatett in his mouþe betokenythe þe bleddyr ys perisshed and yf he /haue in his breste so narowe þat he may onnethe drawe his breþe þat signifyethe þat postym stronge be wexynge of bloode (Hunter 513)

(4b) ¶ Also ʒif þer were melil jpotil in þe mouth. hit be tokenes þe bleddur is percid. ¶ And ʒif he haue þe breste so narow þat he may vnnethʒ draw wynth. hit signifieʒ empostym to be stronge be waxinge of blod. (Cambridge Dd VI 29)

(5a) /¶/ Also yf a man haue me chill rotynn fylþe at his mouthe þat sygnyfieth þe mydrem to be perisshed (Hunter 513)

(5b) ¶ Also ȝif he haue mikil glat. and castij mekel roten filth out at his mouth. hit signefies þe midrif to be parsed. (Cambridge Dd VI 29)

(6a) /¶/ Also yf a man haue /euyll Aboue þe breste þat sygnyfyethe bloode to breke (Hunter 513)

(6b) ¶ And ȝif aman haue euil abouth þe breste. hit signifieȝ þe bleddur to broke. (Cambridge Dd VI 29)

(7a) /¶/ Also yf the seke loke dedely and tere his Cloþis as A man þat ys frantik betokenythe he schall die of þat selfe euyll (Hunter 513)

(7b) ¶ And also ȝif he loket hidoslich. and terreth his cloþus as man þat frentikhit bitakyns þat he schal dye þe same day./ (Cambridge Dd VI 29)

Despite the parallelisms, several prognostic sentences in Hunter 513 have no counterpart in any of the other versions. Furthermore, they do not deal with skin eruptions anymore. As an example, signs numbered seventeen to twenty in the manuscript show no coincidence with any other predictions in the other witnesses:

(8) /¶/ Also yf þe erynn of the seke be colde and his teþe Cold and þe typpe of his nose and his Chynne hange dunward he schall dye with in v. days

(9) /¶/ Also yf the seke turne ofte to þe wall ward and rubbe ofte his nose / thyrles betokenythe þe dethe to be nyghe

(10) Also yf the seke slepe and his mouþe opyn and gapyng vpward aske hym yf he haue euyll in þe wombe of fretynge and yf he caste noughte or he do wepe with þe ryght eye in þe iij day he schall dye

(11) ¶ Also yf the seke turne his fete there his hede laye it sygnyfiethe dethe

Before concluding, it is relevant to mention that Kuhne (1989a: 12) associates the Arabic texts discussed by her with a pseudo-Galenic prognostic text, *De morte subitanea*, clearly linked to the *Capsula eburnea*. She claims that, among the twenty-five prognostic sentences analysed by her, there are some similar ones that are easily identifiable, others with several common elements and the rest with some shared details that point to a common remote relationship (1989a: 13). It is likely that the Hunter 513 and Cambridge Dd VI 29 scribes had different extant originals from the one used by the copyist of the Additional 34111, even for the second version. Thus, both Hunter 513 and Cambridge Dd VI 29 scribes may have used another, possibly Latin, translation.

4. Conclusions

In the present article we have shed light on the distinction between the different pseudo-Hippocratic prognostic treatises written in the Middle Ages, whose aim was to let the physicians and medical practitioners learn about the signs that would predict the imminent death of the patient, if some specific symptoms were present. Among them, the *Capsula eburnea*, a text allegedly written in Greek in the fourth or fifth century, which was translated into Latin between the fifth and the seven centuries, has been the focus of this article. By the seventh century, the text was also translated into Arabic and at the end of the twelfth century retranslated into Latin. These two Latin versions circulated widely in Europe as well as a third Latin version in the late fifteenth century giving rise to translations in vernacular languages in the late Middle Ages. Subsequently, the treatise is found in Middle English.

Being short medical pieces, pseudo-Hippocratic tracts were frequently inserted into other works and, as a result, they may have been overlooked thus-far. This also explains why they have remained comparatively unknown, and the only way to identify parallel copies is by consulting different catalogues and published reference works, and by checking the original manuscripts. An important hindrance is the fact that even specialised catalogues are rarely comprehensive and often do not include cross-references to other catalogues, which makes the identification of parallel texts an arduous task and, consequently, their editing and study. The second obstacle to overcome is the fact that the treatise under consideration appears associated to or receives a wide variety of titles, and it is attributed within its title not only to Hippocrates, but also to other well-known physicians. The fact that they occur under different names in catalogues makes the information they provide sometimes inaccurate.

The absence of a list containing the Middle English witnesses of the *Capsula eburnea* has resulted in our attempt to obtain as many copies of the text as possible in order to narrow the search and finally to establish a reliable collection of manuscripts containing it. In this article, four versions have been examined: GUL Hunter 513; BL Additional 3411, version 1 and version 2, and

CUL Dd VI 29. These Middle English treatises have been compared with three Latin translations, as published by Sudhoff (1915/16) and Sigerist (1921). None of the Middle English manuscripts can be said to be an exact copy of any of the Latin versions, though some ascriptions can be done: Several variants of the different Latin traditions must have been in circulation. Thus, Additional 34111 version 1 clearly follows the Arabic text translated into Latin by Gerard of Cremona, which corresponds to Sudhoff´s second recension. The ascription of the other three copies is not so obvious. At some points, Additional 34111 version 2 has much in common with the first Latin translation, while Hunter 513 and Cambridge Dd VI 29 versions, although following Sudhoff's first recension and Sigerist's transcription in many excerpts, also show passages absent in all the other versions, which makes it difficult to ascribe them completely to any of the explored traditions. All in all, Hunter 513 and Cambridge Dd VI 29 clearly share a large number of features, even if none can be said to be a copy of the other. They might be witnesses of the third thus-far unexplored Latin tradition in its Middle English translations.

To conclude, some aspects still need further research. For instance, there may be extant copies of the *Capsula eburnea* in Middle English that have not yet been identified. Thus, a detailed inspection and continued search among Middle English medical manuscripts could uncover further texts. The editing and analysis of them would be essential to establish their link to the above-mentioned traditions. Similarly requisite is a comparison of the different Middle English versions with translations in other vernacular traditions in order to investigate the circulation and transmission of the original Latin texts.

REFERENCES

Sources

Glasgow University Library Catalogue
https://www.gla.ac.uk/myglasgow/archivespecialcollections/, accessed May 2017

Middle English Dictionary Online (*MED* online)
https://quod.lib.umich.edu/m/middle-english-dictionary/dictionary, accessed April 2020

Oxford English Dictionary Online (*OED* online)
www.oed.com, accessed April 2020

Sloane Manuscripts (*British Library* online)
http://hviewer.bl.uk/IamsHViewer/Default.aspx?mdark=ark:/81055/vdc_100000000040.0x00011c, accessed November 2021

Special studies

Alvar Ezquerra, C.
2001 "Textos científicos traducidos al castellano durante la Edad Media". In: N. Henrad – P. Moreno – M. Thiry-Stassin (eds.) *Convergences medievales. Epopée, lyrique, roman: mélanges offerts a Madeleine Tyssens*. Louvain-la-Neuve: De Boeck Université, 25-47.

Arrizabalaga, J.
1999 "Medical causes of death in preindustrial Europe: Some historiographic considerations", *Journal of the History of Medicine* 54, 241-260.

Beaujouan, G.
1972 "Manuscrits médicaux du Moyen Âge conservés en Espagne", *Mélanges de la Casa de Velázquez* 8, 161-221.

Benati, C.
2013 "The ever-lasting rules of death? The reception and adaptation of the pseudo-Hippocratic *Capsula Eburnea* in German medical literature", *Brathair* 13 (1), 5-18.

De la Cruz Cabanillas, I. – I. Diego Rodríguez
2018 "Medical astrology in Middle English: The case of *Þe Booke of Ypocras*". In: M.J. Esteve Ramos – J.R. Prado-Pérez (eds.) *Textual Reception and Cultural Debate in Medieval English Studies*. Newcastle-upon-Tyne: Cambridge Scholars Publishing, 79-99.

Di Clemente, V.
2011 "Vicende della letteratura medico-prognostica pseudoippocratea nell'Europa medievale: la cosiddetta *Capsula Eburnea* (*Analogium Hippocratis*, *Liber Veritatis Hippocratis*, *Secreta Hippocratis*, *Secreta Democriti*) e la sua ricezione in area alto-tedesca (XI/XII-XV)", *Itinerari* 2, 49-74.
2019 "La tradizione della Capsula eburnea in inglese medio: il caso della doppia versione del manoscritto Londra, British Library, Add. 34111". In: NUME (Gruppo di ricerca sul Medioevo Latino) *V ciclo di studi medievali: Atti del Convegno*, 3-4 giugno 2019. Lesmo: EBS Edizioni, 557-562.

Hunt, T.
1990 *Popular Medicine in the Thirteenth-Century England*. Cambridge: D. S. Brewer.

Keiser, G.R.
1998 *XXV. Works of Science and Education*. In: A.E. Hartung (ed.) *A Manual of the Writings in Middle English, 1050-1500*. Vol. 10. New Haven, CT: Connecticut Academy of Arts and Science, 3593-3967.

Kibre, P.

1945 "Hippocratic writings in the Middle Ages", *Bulletin of the History of Medicine* 18, 371-412.

1977 "Hippocrates Latinus: Repertorium of Hippocratic writings in the Latin Middle Ages (III)", *Traditio* 33, 253-295.

1978 "Hippocrates Latinus: Repertorium of Hippocratic writings in the Latin Middle Ages (IV)", *Traditio* 34, 196-226.

1982 "Hippocrates Latinus: Repertorium of Hippocratic writings in the Latin Middle Ages (VIII)", *Traditio* 38, 165-192.

Kuhne Brabant, R.

1984/85 "El eslabón árabe en la transmisión de los *Secreta Hippocratis*", *Awraq: Estudios sobre el mundo árabe e islámico contemporáneo* 7-8, 31-37.

1986 "Una versión aljamiada del *Secreto de Hipócrates*", *Sefarad* 46 (1), 253-269.

1987/88 "The Arabic prototype of the *Capsula Eburnea*", *Quaderni di Studi Arabi* 5-6, 431-441.

1989a "El Kitab al-dury, Prototipo árabe de la *Capsula Eburnea* y representante más genuino de la tradición de los Secreta Hippocratis (I)", *Al-Qantara* 10 (1), 3-20.

1989b "El Kitab al-dury, Prototipo árabe de la *Capsula Eburnea* y representante más genuino de la tradición de los Secreta Hippocratis (II)", *Al-Qantara* 10 (2), 299-327.

1990 "El Kitab al-dury, Prototipo árabe de la *Capsula Eburnea* y representante más genuino de la tradición de los Secreta Hippocratis (III)", *Al-Qantara* 11 (1), 3-58.

Meyer, P.

1903 "Les manuscrits français de Cambridge. III. Trinity College", *Romania* 32 (125), 18-120.

Muschel, J.

1932 "Todesprognostik und die *Capsula eburnea* in hebräischer Überlieferug", *Sudhoffs Archiv für Geschichte der Medizin* 25 (1), 43-60.

Paxton, F.S.

1993 "*Signa Mortifera*: Death and prognostication in early medieval monastic medicine", *Bulletin of the History of Medicine* 67 (4), 631-650.

Robbins, R.H.

1970 "Signs of death in Middle English", *Mediaeval Studies* 32, 282- 298.

Sigerist, H.E.

1921 "Di Prognostica Democriti im Cod. Hunterian T. 4. 13, S. IX/X", *Archiv für Geschichte der Medizin* 13 (5-6), 157-159.

Sudhoff, K.

1914 „Die kurze „Vita" und das Verzeichnis der Arbeiten Gerhards von Cremona, von seinen Schülern und Studiengenossen kurz nach dem Tode des Meisters (1187) zu Toledo verabfaßt", *Archiv für Geschichte der Medizin* 8 (2-3), 73-82.

1915/16 “Die pseudohippokratische Krankheitsprognostik nach dem Auftreten von Hautausschlägen, „Secreta Hippocratis“ oder „Capsula eburnea“ benannt”, *Archiv für Geschichte der Medizin* 9 (1-2), 79-116.

Tavormina, M.T.
2007 “The Middle English *Letter of Ipocras*”, *English Studies* 88 (6), 632-652.

Voigts, L.E. – P.D. Kurtz
2000 *Scientific and Medical Writings in Old and Middle English: An Electronic Reference*. Ann Arbor: University of Michigan Press.

Young, J. – P.H. Aitken
1908 *A Catalogue of the Manuscripts in the Library of the Hunterian Museum in the University of Glasgow*. Glasgow: MacLehose.

Address: Isabel de la Cruz-Cabanillas, Universidad de Alcalá, Departamento de Filología Moderna, Facultad de Filosofía y Letras, C/ Trinidad, 3, 28801 Alcalá de Henares, Spain.
ORCID code: https://orcid.org/0000-0001-7323-0796

Address: Irene Diego-Rodríguez, Universidad Antonio de Nebrija, Departamento de Lenguas Aplicadas, Facultad de Lenguas y Educación, C/ Santa Cruz de Marcenado, 27, 28015, Madrid, Spain.
ORCID code: https://orcid.org/0000-0002-9608-1436

Token: A Journal of English Linguistics 13, 2021

The representation of the concept of flirtation and coquetry in English: An analysis based on the *Historical Thesaurus of the Oxford English Dictionary*

Julia Landmann

University of Heidelberg

ABSTRACT

Love is as old as humanity itself. Therefore, it is not surprising that over the centuries, a great variety of words and expressions related to this subject have been coined. The *Historical Thesaurus of the Oxford English Dictionary* (henceforth referred to as the *HTOED*) is a rich resource for those who intend to study this field from a linguistic point of view. The present analysis sets out to examine an essential domain which is related to the field of love. It will examine a comprehensive sample of 79 nouns referring to flirtation and coquetry. This has so far been neglected in current research. Besides the *HTOED*, media such as the *OED Online* and corpora of present-day English including the *BNC*, the *COCA*, the *Movie Corpus*, the *TV Corpus* and the *Soap Corpus* will be employed, in order to get a rounded picture of the etymological origin, semantics and contextual uses, including informal usage, of the sample of words under scrutiny from a historical perspective. It thus goes far beyond the scope of previous research on this area.

Keywords: lexicology, online dictionaries and corpora in lexicological research, the *Historical Thesaurus of the Oxford English Dictionary*, vocabulary related to flirtation and coquetry.

1. Studies based on the *Historical Thesaurus of English*

By way of introduction, mention should be made of some book-length studies which rely, to a greater or lesser extent, on data collected from the *Historical Thesaurus*: Coleman (1999), Tissari (2003) and Crystal (2014). On the basis of the linguistic data belonging to the archives of the Glasgow *Historical Theasurus*

DOI: 10.25951/4819

and additional linguistic documentary evidence collected from various sources such as dictionaries, newspapers, books and films, Coleman (1999) has compiled a huge collection of lexical items which have been divided into the semantic fields of love, sex and marriage, from their earliest attested uses in English until recent decades. The findings presented in Coleman's study yield a thesaurus on its own in which the words from the afore-mentioned areas are listed in chronological order. Her study is illuminating in many ways. Yet, Coleman does not further examine the use of the various lexical items under review. For example, an analysis of the contextual use of the terms in English by means of supporting linguistic documentary evidence in dictionaries such as the *OED* or corpora would have been desirable.

Tissari's study from 2003 entitled *LOVEscapes: Changes in Prototypical Senses and Cognitive Metaphors since 1500* includes one chapter which presents findings resulting from the analysis of the *Historical Thesaurus*, which was still in the production phase at the time of her research. Tissari used some data from the thesaurus in chapter fourteen about "Love in Words", in order to interpret her general findings with respect to the semantics of the word *love* against the historical background and development of the domain of *love*.

In his book *Words in Time and Place* (2014), Crystal selects fifteen semantic domains including words related to pop music, privies, oaths, and other areas. A brief section of his study is dedicated to vocabulary related to love, i.e. to terms of endearment (Crystal 2014: 103-116). He investigates the development and change of the English language over the centuries in these domains, evaluating the linguistic information which is provided by the *HTOED*. Crystal offers possible reasons why new words have occurred in English in the relevant semantic areas. In addition, he explores the etymological origin of the lexical items in the various domains and addresses the question of how they reflect the socio-cultural background of the period in which they are first recorded.

A number of articles are based on research into specific semantic domains carried out by means of the *HTOED*. Examples are Roberts (2002), O'Hare (2004), Sylvester (2006), Wild (2010), Díaz Vera (2012), Newman (2013), Allan (2015), Roberts and Sylvester (2017). The focus of Roberts's 2002 paper is on the representation of early Middle English in the *Historical Thesaurus of English*. Roberts offers illustrative linguistic examples from this period, including vocabulary from the field of war and peace. O'Hare's article from 2004 is concerned with folk classifications which can be found in the taxonomic system of lexical items referring to plants in the *Historical Thesaurus of English*, and Sylvester (2006) examines the question of whether and to what extent social and moral attitudes are reflected by the conceptual categorization of the *Historical Thesaurus*. The focus of her study is on lexical items referring to consent, coercion and resistance in association with what she refers to as "conceptualizations of sexual contracts across time" (Sylvester 2006: 186). Wild (2010) concentrates on vocabulary from the semantic domain of childhood, while Díaz Vera (2012) investigates the metaphorical conceptualizations of jealousy in Shakespeare's plays, resulting from a close perusal of all the terms for jealousy which are recorded from 1500 to 1700 in the *HTOED*. Newman (2013) explores words denoting 'sailor' from Old English times down to the present. Allan (2015) looks at the vocabulary of education in the *HTOED*, which was subjected to essential changes over the centuries. She addresses the important issue of whether and to what extent these changes reflect different understandings of the concept itself in the history of English, and how it is associated with other related conceptual domains, such as teaching and learning. Roberts and Sylvester (2017) focus on the vocabulary of error and analyse its usage down the ages by considering the development of English orthography, its pronunciation and grammar. By means of the linguistic data provided by the *Historical Thesaurus*, the authors illustrate how notions of errors and mistakes have changed throughout the history of the English language.

Various studies concentrate on the analysis of borrowed lexical items on the basis of the linguistic documentary evidence recorded in the *HTOED*. For example, Coleman (1995) makes a comparison between borrowings from Latin and French in several different semantic fields (i.e. love, sex, hate, marriage) from Old English times to the close of the twentieth century. Her survey relies on a careful perusal of the linguistic data found in the Glasgow *Historical Thesaurus*. The number of borrowings that were investigated by Coleman amounts to 11,000 lexical items. Of these, 1719 were grouped into the area of hate, 3157 in love, 3067 in sex and 1370 in marriage (see Coleman 1995: 102). In addition, Coleman looks at the semantic and morphological evolution of the borrowed lexical items, reflecting on the extent to which they have been assimilated into English. Durkin's (2016) article should also be mentioned here. It addresses both potential advantages and issues that may arise in connection with the use of resources such as the *OED* and the *HTOED*. Durkin's paper sets out to discover how these sources can be used in combination with others to give enlightening results associated with borrowing processes and their effects on the core vocabulary of the English language from a historical perspective. The work utilizing the *HTOED* in combination with dictionaries such as the *OED* is illustrated by a number of examples of borrowed lexical items, encompassing words of French origin (e.g. *carry*, *cry* and *soil*) (see also Durkin 2016: 393). Sylvester (2018) examines a proportion of technical vocabulary borrowed from French during the Middle English period. More specifically, she looks at French-derived terms and their native equivalents referring to instruments within the semantic domain of building used in mediaeval times. The information retrieved from the *Middle English Dictionary* is complemented by lexical items included in other sources, such as the *Historical Thesaurus*.

2. Aims of the present study

As pointed out before, the focus of this study is on a lexical-semantic domain which is related to the superordinate field of love. That is, the sample of nouns that represent essential concepts and motifs related to flirtation and coquetry.

More than a mere count of the flirtation and coquetry terms, the present investigation will offer a detailed analysis of their etymology, meaning and contextual usage in English (encompassing informal language), which has as yet figured little if at all in previous surveys. An important aim of this study is to identify those lexical items which have become comparatively widespread in English. Terms which have made it into common usage will be contrasted with words which have become rare or obsolete in present-day English.

A historical perspective will be assumed to determine the context in which a word has been embedded since its earliest recorded usage. A contextual evaluation can provide essential clues with respect to the reasons why certain lexical items have become widespread in English, while others are confined to specific contexts or have become disused. This also raises the question of which terms occur in colloquial usage or slang. Words which are restricted to particular regional or national varieties of English will also be identified in the present investigation.

Furthermore, an overview of the chronological distribution of the various lexical items throughout the centuries will be given. The question of whether the numbers and proportions of the words under scrutiny is constant or changing over time will be addressed. The present study sets out to assess and provide reasons for variability between speakers and their selection of the different words under scrutiny down the ages. The senses and usages of words have to be studied by assuming a historical point of view, considering trends and developments in the language and culture of the corresponding language community which may explain variation in linguistic usage. Such a comprehensive analysis of the terms relating to flirtation and coquetry is missing in existing research.

3. Methodology

The following section offers an overview of the methodology which was developed for the present investigation. As stated above, the *HTOED* provided the data for the present analysis. Under the direction of a team of scholars, including Christian Kay, Jane Roberts and Irené Wotherspoon, the *HTOED* had been completed at the University of Glasgow over a time span of 45 years before it was released by Oxford University Press in 2009 (see also Kay et al. 2001: 173). The *HTOED* divides meanings and lexical items given in the *OED* on the basis of their subject field, and arranges them by their earliest recorded usage in English. It serves as a taxonomically arranged network of words and senses reflecting the development and history of the English language. The *HTOED* enables its users to examine the variety of lexical items used for a specific sense or concept over the centuries, which makes it a unique source that goes far beyond a typical thesaurus, such as Roget's (2004 [1852]) *Thesaurus of English Words and Phrases*, which constitutes one of the earliest thesauri of English, or studies focusing on a collection of lexical items from a selection of semantic domains (e.g. Coleman's study of words from the fields of love, sex and marriage from 1999), or thesauri concentrating on language use of individual authors (e.g. Spevack's *A Shakespeare Thesaurus* from 1993) which were written in more recent decades.

The *HTOED* was initially based in part on information found in Roget's thesaurus with the corresponding approach to classification, and was then extended with the linguistic documentary evidence retrieved from the new electronic form of the *OED*. In addition, electronic text corpora were also used to determine the semantics of a lexical item and to check if it was still current.

The *HTOED* comprises two volumes. Volume one represents the thesaurus, while volume two comprises an index of the majority of lexical entries found in the thesaurus. It seems noteworthy that volume two excludes items which are only documented in Old English and phrases which encompass more than four words. The printed version and its electronic form that is searchable via the *OED Online* make it possible to look for lexical items and semantic domains, despite the fact that lexical items from Old English are not included in the online edition. The electronic version is directly linked to the information provided by the *OED* and allows, for instance, for an evaluation of the entire semantic scope of a given lexical entry.

The *HTOED* represents the first comprehensive thesaurus to categorize lexical items on the basis of their meaning and according to their earliest attested usage. The linguistic data of the *Historical Thesaurus* initially included about 650,000 slips of paper retrieved from the documentary evidence used for the *OED*, its supplementary volumes and dictionaries of Anglo-Saxon. These paper slips were originally arranged in line with the semantic classification found in Roget's *Thesaurus of English Words and Phrases*. Yet, the team of the *Historical Thesaurus* Project decided to re-classify the linguistic material, in order to deal with such a large lexicological sample (see also Kay 1994: 67).

The categorization now reflects a variety of overriding semantic domains, including the external world, the mind, and society as three major categories which themselves comprise a multitude of subfields. The lexical items within these areas are divided into semantic hierarchies and arranged chronologically according to their earliest attested use in English. The categorization of the lexical items in the *HTOED* represents a frame of reference for the classification carried out in the present study. The various flirtation and coquetry terms were collected from the *HTOED* in the spring of 2020 (for details on the procedure to identify these words, see section 4 below).

The *HTOED* reflects the semantic development of the English lexicon in all its variety, since it documents usages from the Old English period until today. In total, it comprises nearly 800,000 words and senses, providing a comprehensive collection of meaning categorized according to semantic fields. It mainly relies on the second edition of the *OED* published in 1989 as well as the *Thesaurus of Old English* (see also <https://ht.ac.uk/>). By means of the *HTOED*, synonyms for particular lexical entries recorded in the *OED Online* can be identified. In addition, their sense development from their earliest attested usage up to the present day can be assessed due to the rich linguistic documentary evidence in the *OED Online*.

As to the *OED*, it is currently being revised. The entire text of the *OED* can be searched online at <http://www.oed.com>, where the linguistic data is being updated every quarter with the preliminary findings of the *OED* rewrite. The *OED Online* includes the complete text of the Second Edition of the dictionary (henceforth, *OED*2), the two supplementary volumes of the *OED Additions Series* from 1993 and 1997, and a considerable number of updated and new lexical entries which belong to the planned *OED3*.

To research the etymology, the semantics and usage of the sample of words collected from the *HTOED*, the linguistic information included in the *OED Online* was equally taken into account. The documentary evidence in the *OED* allows for an analysis of the typical use(s) in which a word is documented in English.

The reader should observe that *OED* entries which have not yet been updated do not comprehensively reflect the entire semantic scope of a word with its different contextual usages from its earliest attestation down to the present day. Hence, additional linguistic material was collected by means of corpora representative of present-day English (e.g. the *BNC*, the *COCA*), in order to find more of the supporting documentary evidence on the basis of which the usage and semantics of a lexical item can be analysed. Only those entries from the *HTOED* list for which there were no recent usage examples available in the unrevised edition of the *OED* were checked. All the corpora in question were consulted to find typical uses of these words in today's English (including informal language).

The search facilities in the *BNC* and the *COCA* make it possible to examine the contextual use of a word in various genres and registers. The *BNC*, initially compiled by Oxford University Press in the 1980s and the 1990s, comprises 100 million words of text reflecting British English towards the close of the twentieth century. It includes a broad range of genres, such as newspapers, fiction, academic literature and spoken data. The *COCA* comprehensively documents American English usage. At present, it includes 560 million words retrieved from divers sources, such as newspapers, fiction, academic texts and spoken material. It records language use from 1990 to 2017.

As will be seen, a number of the words under review are colloquial or slang terms that occur pre-dominantly in informal usage. In addition to the *BNC* and the *COCA*, which provide a rounded picture of the use of a lexical item in present-day English including several different registers, corpora revealing informal language (i.e. the *Movie Corpus*, the *TV Corpus* and the *Soap Corpus*) were consulted. The *Movie Corpus* encompasses 200 million words from more than 25,000 films produced from the 1930s until today, while the *TV Corpus* includes 325 million words of text from 75,000 television episodes between 1950s and the present day. By means of these two resources, the typical informal usage a word shows in English (and in different national varieties of English) can be identified. The *Soap Corpus* enables its users to search for informal usages in American English.

It consists of 100 million words found in 22,000 transcripts of American soap operas of the early twenty-first century.

In order to identify those items which occur comparatively frequently in present-day English, EFL (*English Foreign Language*) dictionaries such as the recent editions of the *OALD* (available online at <https://www.oxfordlearnersdictionaries.com/>) and the *LDOCE* (accessible online at <https://www.ldoceonline.com/>) were used since these dictionaries record those lexical items which are assumed to be known to the "average" native speaker of English. The relatively common flirtation and coquetry terms from the *HTOED* list which could also be found in the EFL dictionaries consulted were contrasted with those which were recorded merely once in English (e.g. *whiting's eye*).

With regard to the questions of which words belong to colloquial use, which terms are used in slang, and which lexical items are rare or extinct, the labelling in the *OED* was drawn on.

4. Terms related to flirtation or coquetry in the *HTOED*

The taxonomic categorization of the terms referring to flirtation or coquetry in the *HTOED* is: "the mind > emotion > love > flirtation or coquetry." The number of terms related to flirtation and coquetry which are part of the *HTOED* amounts to 169 lexical items. Of these, 79 are nouns. These nouns have been examined in detail in the present paper. The following diagram reflects their chronological distribution down the ages:

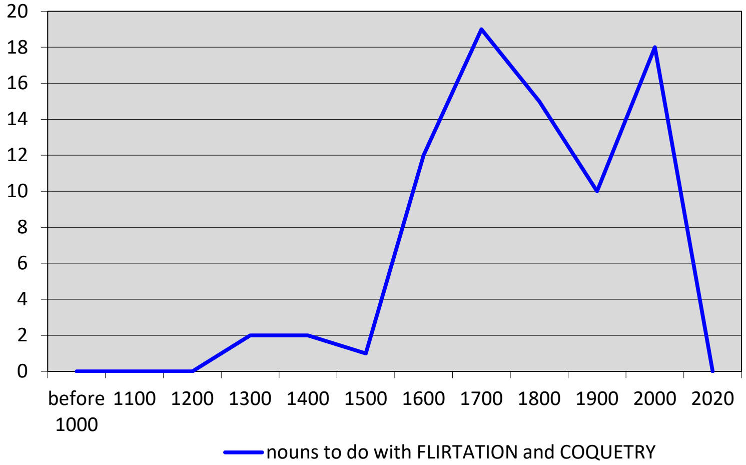


Table 1. The chronological distribution of the nouns related to flirtation and coquetry

According to *HTOED*, there was hardly any use of nouns referring to flirtation and coquetry in English before 1500. From 1501, the number increased to its peak in the seventeenth century. After that, it decreased again until 1901. In the 20th century there was a further increase. The field of flirtation and coquetry does not include any lexical items that date from the twenty-first century.

The list below reflects the numbers and proportions of the nouns related to flirtation and coquetry over the centuries. For each century, the corresponding vocabulary collected from the *HTOED* has been provided. It should be noted that the following symbols have been placed after a word's first recorded use:

∞	confined to colloquial usage or slang
•	rare
†	obsolete

Of the 79 nouns related to flirtation and coquetry, 10 words (12.7%) are confined to colloquial usage or slang, 6 lexical items (i.e. 7.6%) have become rare, and 19 terms (i.e. 24.1%) have become obsolete. Some 12 words (i.e. 15.2%) are recorded in EFL dictionaries such as the *OALD* and the *LDOCE* and can therefore be classified as relatively frequent terms in present-day English. These types of words are printed in bold in the list below.

a) 1201-1300

love-late (might have been first documented in English in circa 1225)†; *tugging* (might have been first recorded in English in circa 1225)†.

b) 1301-1400

love-lake (circa 1330)†; ***dalliance*** (circa 1385).

c) 1401-1500

gallant (about 1450).

d) 1501-1600

simper-de-cocket (about 1529)†; *toy* (1565)†; *toying* (1565-1573); *love-trick* (1567); *dallier* (1568); *amorets* (1590)†; *belgard* (1590)†; *oeillade* (1592); *amorist* (about 1595); *woman's man* (1597); *love-sport* (1598)†; *lady-monger* (circa 1600)†.

e) 1601-1700

sheep's eye (1604); *encounterer* (1609)†; *belamour* (1610)†; ***coquette*** (1611); *lapling* (1628)†; *spider-caul* (1631)†; *dammaret* (1635)†; *rover* (1638); *amoretto* (1647)†; *pickeering* (1650); ***coquetry*** (1656); *gallanting* (1664); *ogle* (about 1668); *whiting's eye* (1673)†; *languishment* (1676); *philander* (1676)•; *ogling* (1682); *coquetting* (1690); *coquet* (1693).

f) 1701-1800

topgallant (1701)†; *flirting* (1710); *languisher* (1713); *toyer* (about 1713); ***flirtation*** (1718); *dangler* (1728); ***flirt*** (about 1732); *agapet* (1736)†; ***philandering*** (1737); *minauderie* (1763)†; ***ladies' man*** (1764); *male coquette* (1770); *Jack among the maids* (1785)•; *agacerie* (1787)•; *ladykilling* (1795).

g) 1801-1900

Lochinvar (1811); *flirter* (1814); *dead set* (1823); ***philanderer*** (1841); *flirtee* (1862); *coquettishness* (1872); *garrison-hack* (1876)∞; *allumeuse* (circa 1891); *philander* (1898)•; *poodle-faker* (1900)∞.

h) 1901-2000

poodle-faking (1902)∞; *Romeo* (1902); *vampire* (1903); *play* (1905)∞; *carryings-on* (1909); *monkey parade* (1910)•∞; ***the glad eye*** (1911)∞; ***vamp*** (about 1911); ***lounge lizard*** (1918)∞; *vamping* (1918); *tea-hound* (1921)•∞; *skirt duty* (1922)∞; *vampishness* (1922); *glad* (1927); *vampiness* (1928); *monkey parading* (1934)∞; *lizard* (1935); *kikay* (1993).

The vocabulary related to flirtation and coquetry is characterized by variety. Besides *coquetry*, which was adapted from the French *coquetterie*, and *flirtation*, the *HTOED* includes the derivations *coquettishness*, *coquetting* and *flirting*. Additional terms referring to flirtatious action and behaviour are *gallanting*, which can be paraphrased as 'flirting', 'giving polite attention to a woman', the common term *ladykilling* and the plural form *carryings-on*, which has been recorded since 1919 in the meaning of "questionable or *outré* proceedings, flirtations, frolics" (*OED2*). As to *agacerie*, a French-derived term for "allurement, coquetry" (*OED3*), the word has become rare in English. The latest *OED3* example dates from 1919:

(1) "1919 *Printers' Ink* 30 Jan. 126/2 Practice all the agacerie known to the advertising art."

Minauderie, a French-derived expression relating to flirtatious behaviour and coquetry, is no longer documented in current usage. There is also the phrase *skirt duty*, a metaphorical term for behaviour aimed at getting to know men. It can also refer to a feeling of obligation to bear women company.

It is important to note that especially the earlier terms which belong to this area reflect the pre-dominant concept of a flirtation and coquetry as a game or trick, as sport and fun. Examples of terms which date from the thirteenth century are *tugging*, literally 'dragging (sportively)', which specifies the action of struggling in an amorous manner, and *love-lake* (now disused), which has the same meaning as *love-sport*. *Dalliance* also falls into this category of lexical item. It was already attested in English in 1385, and has since become quite a widespread term for "[s]port, play (with a companion or companions); esp[ecially] amorous toying or caressing, flirtation; often, in bad sense, wanton toying" (*OED*2). Examples from the sixteenth century comprise *toy*, *toying*, *dallier*, *love-trick* and *love-sport* (now obsolete). *Pickeering*, "[w]ordy, playful, or amorous skirmishing; wrangling, bickering, petty quarrelling; an instance of this" (*OED3*), is an example from the seventeenth century. The term has become archaic in current usage. It is now confined to regional, i.e. Irish English, as in:

(2) "1998 *Irish Times* (Nexis) 24 Oct. 28 Pickeering, if I may refresh your memory, is the act of making romantic overtures to a woman. It was, he says, much used in his own boyhood in Co Westmeath." (*OED3*)

Play has been documented since 1905 in a specialized meaning, one referring to "an attempt to sexually attract another person" (*OED3*). In this use, it frequently occurs in the verbal phrase *to make a play* (*for*), e.g.:

(3) "You don't think! I make a play for every lady who comes by here, do you? You got chemistry." (*Microwave Massacre*, 1983, *Movie Corpus*)

(4) "In this one other people besides his daughter try to fix him up with women, but it never works out, mainly because neither is interested in most of the things the other is. In this one he doesn't make a play for women much younger than he." (*Harper's Magazine*, 2016, *COCA*)

The collection of words referring to flirtation and coquetry also comprises several terms which specify different expressions of the eyes, as *pars pro toto* for the person who flirts. Some terms specify flirtatious glances, ranging from the metaphorical terms *the glad eye*, *sheep's eye* and *whiting's eye*, to *ogle*, *ogling*, *oeillade*, *love-late*, *amorets*, *belamour*, *amoretto* and *belgard*. Of these, *oeillade*, a borrowing from French, specifies a flirtatious glance as a sign of affection or desire, "an ogle" (*OED3*). *Love-late*, referring to an amorous glance or flirtatious behaviour, has become obsolete in English. The latest *OED3* quotation is from about 1400:

(5) "*a*1400 *Ancrene Riwle* (Pepys) (1976) 38 Ʒiue me þi louelates, ȝe, to me and to non oþer."

Just like *love-late*, the French borrowings *amorets* and *belamour*, *belgard*, an Italian-derived term, and *amoretto*, which might be an alteration of *amoret*, are no longer documented in current usage. All these terms specify looks of love. Similarly, *whiting's eye*, an affectionate or sensual glance, is now obsolete in English. As to *the glad eye* (also abbreviated as *glad*), it is also recorded in the idiomatic expression *to give someone the glad eye*, now an archaic phrase common in British English in the meaning of 'to ogle.' *Languisher* refers to an individual assuming a specific glance, i.e. to someone who adopts a listless expression or a person who has a loving or longing gaze. The corresponding abstract term *languishment* can also be found among the *HTOED* entries to do with expressions of affection or aspiration. The reader might observe that most of the terms specifying looks of love date from the sixteenth and seventeenth centuries. The only example from the twentieth century is *the glad eye* (also abbreviated as *glad*). One may thus conclude that this perspective or understanding of flirtation and coquetry attaching specific value to the expression of the eyes seems to have lost importance among speakers of present-day English.

Another typical feature of this group of words is the considerable number of names for persons engaged in flirtation and coquetry that it takes in, among them a number of pejorative terms chiefly occurring in informal usage or slang. The number of colloquial terms has increased since the nineteenth century. The following terms relate to men engaged in flirtation and coquetry: *rover*, *flirt*, *poodle-faker*, *lounge lizard*, *tea-hound*, *coquet*, *male coquette*, *gallant*, *topgallant*, *dallier*, *dangler*, *amorist*, *woman's man*, *ladies' man*, *lady-monger*, *dammaret* and *spider-caul*. Of these, *rover*, generally referring to 'a roamer', 'a vagrant', shows a meaning relating to flirtation which has become historical in present-day English, i.e. it is confined to contexts somehow related to the past. Since 1638, it has been used to specify 'a man who flirts a lot or is not faithful', 'a man who is fickle in his feelings', as in the following example which deals with the theatre play *The Rover*, written by Aphra Behn in the seventeenth century:

(6) "1998 M. Zook in *Women Writers & Early Mod. Brit. Polit. Trad.* iv. 78 The transformation of Behn's cavalier from merry rover to cunning rake to sacrificial martyr-hero." (*OED3*)

As to *flirt*, it has been used with reference to a man engaged in flirting since about 1732 and with respect to a female flirt since 1747 (see *OED2*). The word came to denote the individual one flirts with, i.e. a *flirtee*, in 1779:

(7) "1779 *Gentleman's Mag.* **49** 357 The General [Howe] has found another Desdemona at Philadelphia … who is now his Excellency's flirt." (*OED2*)

Poodle-faker functions as a slang term for "[a] man who cultivates female society, esp[ecially] for the purpose of professional advancement; a ladies' man; an unmanly man" (*OED3*). *Poodle-faking*, the corresponding noun describing the behaviour of such a man can also be identified in the *HTOED*. Typical usage examples from the corpora consulted are:

(8) "Well, don't try to involve me into your sordid poodle-faking! It's not poodle-faking. I love her." (*You Rang*, M'Lord?, 1990, *TV Corpus*)

(9) "Are you sure? No. Chasing a collection of feckless poodle-fakers across three countries! I am an examining magistrate." (*The Abduction Club*, 2002, *Movie Corpus*)

Lounge lizard (also abbreviated as *lizard*) and *tea-hound* represent metaphorical expressions confined to American English slang. The former refers to a man who mingles with the rich instead of working and is looking for a wealthy female partner who can support him financially, as is illustrated by the following examples:

(10) "How can you leave her with that money-sniffing lounge lizard? I mean, where's your sense of friendship?" (*All My Children*, 2004, *Soap Corpus*)

(11) "Over a weekend celebrating the rapper's induction into the Bronx Walk of Fame, Slick Rick swapped his signature eyepatch for bejeweled

Chanel sunglasses and his onerous gold chains for a sequined bow tie, but the outlandish lounge lizard suits could only have been his." (*Vanity Fair*, 2018, *COCA*)

Tea-hound has become a rare term for "a man given to frequenting tea-parties; also in extended use, a lady's man" (*OED*2).

In its function as an intensifying form of *gallant*, the word *topgallant* is disused in current English. *Lady-monger* and *dammaret* (with its spelling variant *damouret*) which serve as synonyms of 'a ladies' man', *lapling*, 'a male person fond of lying on a woman's lap', *agapet*, "[a] lover of women, a philanderer" (*OED*3), and *spider-caul*, literally 'spider's web', a designation of a flirting man, have also become obsolete. The latest *OED* examples illiustrating the use of these terms are:

(12) "1707 E. Ward *London Terraefilius* No. 1. 26 That Libidinous Coxcomb of a Creature, is one of those Insatiate Lady-mongers, call'd an Universal Lover." (*OED*2)

(13) "*a*1649 W. Drummond *Hist. Scotl.* (1655) sig. L2^{v} Place me with a *Damouret* … if I praise him in the presence of his Mistress, he will be ready to perform like duties to me." (*OED*2)

(14) "1658 J. Hewitt *Repentance & Conversion* 7 You must not stream out your Youth in Wine and live a Lapling to the Silk and Dainties." (*OED*2)

(15) "1775 J. Ash *New Dict. Eng. Lang. Agapet*, A lover of the fair sex, a man of pleasure. [Also in later dictionaries.]" (*OED*3)

(16) "1631 R. Brathwait *Engl. Gentlewoman* 93 Let not then these Spider-cauls delude you, discretion will laugh at them, modesty loath them." (*OED*2)

Of these, *dammaret/damouret* and *agapet* constitute borrowings. The former reflects the French *dameret* 'a man dedicated to the entertainment and courting of ladies', and the latter was derived from the Byzantine Greek ἀγαπητός in the meaning of 'man having a secret female sweetheart' (see *OED*3).

The *HTOED* also comprises words and phrases derived from proper nouns which are used to specify male flirts and coquets, such as *Romeo*, *Lochinvar* and *Jack among the maids*. *Romeo* reflects the name of the protagonist in Shakespeare's play *Romeo and Juliet*. It has become a widespread term in English for a "[a] seducer or habitual pursuer of women; a philanderer, a womanizer" (*OED3*), as in:

(17) "In a strange way it can actually contribute to the success of many marriages, for not every woman wants to spend her life with an eternal Romeo." (*Caring for Elderly Parents*, 1979, *BNC*)

(18) "Bill: No, wait, just cool your jets, Romeo. I'm just trying to look at the situation objectively." (*Bold and Beautiful*, 2012, *Soap Corpus*)

Lochinvar corresponds to the name of the chief character of a ballad in *Marmion* written by the Scottish poet Sir Walter Scott. He runs away with a woman on the day of her wedding, in order to marry her secretly. In present-day English, *Lochinvar* specifies "[a] male eloper; a heroic or daring male lover" (*OED3*). The expression *Jack among the maids* is now a rare term for 'a man who likes to surround himself with women', 'a woman's favorite.' It may also function as a synonymous idiom for 'the cock of the walk.'

Several terms in the *HTOED* entries relate to female flirts and coquettes. Examples are *allumeuse*, *garrison-hack* and *simper-de-cocket*. As regards *allumeuse*, the word has its origins in French, and refers to "[a] woman who is alluring but sexually elusive; a flirt, coquette (usually with some degree of sophistication implied)" (*OED3*), e.g.:

(19) "2008 H. Kaiser *French War Brides in Amer.* ii. 12 I remember that one of the war brides was called Paulette Sparks. She was very pretty and lively. She was also a real flirt, an *allumeuse*, and she was married twice." (*OED3*)

The metaphorical expression *garrison-hack* relates to a woman flirting extensively with soldiers of a garrison. The etymological origin of *simper-de-cocket*, "[a]n affected coquettish manner; a woman having such a manner; a flirt" (*OED3*), is not perfectly clear. According to the *OED3*, it might have been formed on the basis of the verb *simper*, "[t]o smile in a silly, self-conscious, or affectedly coy or bashful manner, or in a way that is expressive of or is intended to convey guileless pleasure, childlike innocence, or the like" (*OED3*), and the French *coquette*, but this theory has not yet been proved. The word belongs to the group of lexical items which are no longer documented in present-day usage.

The concept of a seducing, attractive woman considered a vamp occurred in the twentieth century: *vampire* has been recorded in the relevant sense since 1903, and the corresponding abbreviation *vamp* has been documented in English since 1911. The Tagalog borrowing *kikay*, "[a] flirtatious girl or woman", "a girl or woman interested in beauty products and fashion" (*OED3*) represents the latest term in this semantic domain. It was borrowed into English in 1993. According to the *OED3*, it is confined to Philippine English.

5. Summary and conclusion

An important characteristic of the vocabulary related to flirtation and coquetry is its diversity. As was seen in the present study, the *HTOED* encompasses a variety of lexical entries from this domain, ranging from terms specifying flirtatious action and behaviour, to lexical items referring to flirtatious glances and names for persons engaged in flirtation and coquetry.

The *HTOED* includes twelve fairly widespread lexical items which belong to the vocabulary included in EFL dictionaries such as the *OALD* and the *LDOCE*. As mentioned earlier, these dictionaries represent reliable sources for identifying comparatively frequent and well-used words in present-day English. They make up 15.2% of the hyponyms of the vocabulary related to coquetry and flirtation. These are *dalliance* (circa 1385), *gallant* (about 1450), *coquette* (1611), *coquetry* (1656), *flirtation* (1718), *flirt* (about 1732), *philandering* (1737), *ladies' man* (1764), *philanderer* (1841), *the glad eye* (1911), *vamp* (about 1911) and *lounge lizard* (1918). Several items in the category of fairly common words and phrases are more general terms referring to flirtation and coquetry. Examples are the two terms themselves (i.e. *flirtation*, *coquetry*) as well as *dalliance* and *flirt*. It also includes a number of comparatively frequent names for persons involved in flirtation and coquetry, ranging from *gallant* and *coquette*, to more recent expressions such as *vamp* and the slang term *lounge lizard*.

Socio-cultural trends may explain particular tendencies in the development of the vocabulary of a language. In this paper, some overarching trends have been discovered, such as the post-eighteenth-century general increase in colloquial items, terms which are fairly common, especially in corpora reflecting informal language. Of the 79 nouns referring to flirtation and coquetry, ten lexical items (i.e. 12.7%) are identified as words restricted (or chiefly restricted) to colloquial English or slang. These types of lexical items can be found only in the nineteenth and twentieth centuries – with

20 and 80% respectively, i.e. showing a significant increase in the twentieth century. Examples are terms for (typically male) individuals and their behaviour, such as *tea-hound*, *lounge lizard*, *skirt duty* and *poodle-faking*. Yet, the fact that colloquial terms are present only in the nineteenth and twentieth centuries does not necessarily mean that such items did not exist before – it may simply be that the earlier sources rarely exhibit colloquial items and/or terms that are not classified as colloquial in the dictionaries consulted.

The six flirtation and coquetry terms that have become rare in present-day English were first documented in the *OED* in the period from 1601 to 2000. Most of the flirtation and coquetry terms first recorded between 1501 and 1700 have become disused in English.

The present study also identified essential concepts and motifs prevalent in the semantic domain under examination, such as the understanding of flirtation and coquetry attributing particular importance to the expression of the eyes in the sixteenth and seventeenth centuries, and the motif of the *vamp* which occurred in the early twentieth century. Reaching beyond the scope of this investigation, future studies might, desirably, compare additional semantic fields related to love.

REFERENCES

Sources

British National Corpus (*BNC*)
https://www.english-corpora.org/bnc/, accessed July 2020

Corpus of American Soap Operas (*Soap Corpus*)
https://corpus.byu.edu/soap/, accessed July 2020

Corpus of Contemporary American English (*COCA*)
https://www.english-corpora.org/coca/, accessed July 2020

Longman Dictionary of Contemporary English (*LDOCE*)
https://www.ldoceonline.com/, accessed July 2020

Movie Corpus
https://www.english-corpora.org/movies/, accessed July 2020

Oxford Advanced Learner's Dictionary (*OALD*)
https://www.oxfordlearnersdictionaries.com/, accessed July 2020

Oxford English Dictionary Online (*OED online*, including the *Historical Thesaurus*)
http://www.oed.com, accessed July 2020

Thesaurus of Old English
https://ht.ac.uk/, accessed July 2020

TV Corpus
https://www.english-corpora.org/tv/, accessed July 2020

Special studies

Allan, K.
2015 "Education in the *Historical Thesaurus of the Oxford English Dictionary*". In: J. Daems et al. (eds.) *Change of Paradigms – New Paradoxes: Recontextualizing Language and Linguistics*. Berlin; Boston: De Gruyter Mouton, 81-95.

Coleman, J.
1995 "The chronology of French and Latin loan words in English", *Transactions of the Philological Society* 93, 95-124.
1999 *Love, Sex, and Marriage. A Historical Thesaurus*. Amsterdam: Rodopi.

Crystal, D.
2014 *Words in Time and Place. Exploring Language through the* Historical Thesaurus of the Oxford English Dictionary. Oxford: Oxford University Press.

Díaz Vera, J.E.
2012 "Infected affiances. Metaphors of the word JEALOUSY in Shakespeare's plays", *Metaphorik.de* 22, 23-43.

Durkin, P.
2016 "The *OED* and *HTOED* as tools in practical research: A test case examining the impact of loanwords on areas of the core lexicon". In: M. Kytö – P. Pahta (eds.) *The Cambridge Handbook of English Historical Linguistics*. Cambridge: Cambridge University Press, 390-406.

Kay, C.
1994 "Word lists for a changing world". In: W. Hüllen (ed.) *The World in a List of Words*. Tübingen: Niemeyer, 67-75.

Kay, C. – L. Sylvester – I. Wotherspoon
2001 "One thesaurus leads to another". In: C. Kay – L. Sylvester (eds.) *Lexis and Text in Early English: Studies Presented to Jane Roberts*. Amsterdam: Rodopi, 173-186.

Newman, J.G.
2013 "Words denoting 'sailor' in the history of the English lexicon: Abundance, variation, and diction". In: G. Dimković-Telebaković (ed.) *Foreign Language in Transport and Traffic Engineering Profession and Science*. Belgrade: University of Belgrade, Faculty of Transport and Traffic Engineering, 15-29.

O'Hare, C.
2004 "Folk classification in the *HTE* 'Plants' category". In: C. Kay – J.J. Smith (eds.) *Categorization in the History of English*. Amsterdam: John Benjamins, 179-191.

Roberts, J.

2002 "Some thoughts on the representation of Early Middle English in the *Historical Thesaurus of English*", *Dictionaries: Journal of the Dictionary Society of North America* 23, 180-207.

Roberts, J. – L. Sylvester

2017 "Blunder, error, mistake, pitfall: Trawling the *OED* with the help of the *Historical Thesaurus*", *Altre Modernità*, Special Issue: *Errors: Communication and its Discontents* (guest edited by P. Caponi – G. Iamartino – D. Newbold), 18-35.

Roget, P.M.

2004 [1852] *Thesaurus of English Words and Phrases*. Edited by G. Davidson. London: Penguin Books.

Spevack, M.

1993 *A Shakespeare Thesaurus*. Hildesheim; Zürich: Olms.

Sylvester, L.

2006 "Forces of change: Are social and moral attitudes legible in this *Historical Thesaurus* classification?". In: G.D. Caie – H. Hough – I. Wotherspoon (eds.) *The Power of Words. Essays in Lexicography, Lexicology and Semantics. In onour of Christian J. Kay*. Amsterdam; New York: Rodopi, 185-208.

2018 "Contact effects on the technical lexis of Middle English: A semantic hierarchic approach", *English Language and Linguistics* 22 (2), 249-264.

Tissari, H.

2003 *LOVEscapes: Changes in Prototypical Senses and Cognitive Metaphors since 1500*. Helsinki: Société Néophilologique.

Wild, K.

2010 "Angelets, trudgeons, and bratlings: The lexicalization of childhood in the *Historical Thesaurus of the Oxford English Dictionary*". In: M. Adams (ed.) "*Cunning passages, contrived corridors": Unexpected Essays in the History of Lexicography.* Monza: Polimetrica, 289-308.

Address: Julia Landmann, University of Heidelberg, English Department, Kettengasse 12, 69117 Heidelberg, Germany.
ORCID code: orcid.org/0000-0002-2077-2169

A database of early modern first citations from the OED: Religious and geographical terminology[1]

Angela Andreani* and Daniel Russo**

**Università degli Studi di Milano*

***Università degli Studi dell'Insubria*

ABSTRACT

In the wake of the Reformation, intellectuals from all parts of the religious spectrum read, studied and translated Christian sources, not only the Scriptures but also ancient and modern patristic sources, sermons, commentaries, chronicles. The users of these texts – translators, theologians, controversialists – were highly experimental and lexically innovative, as demonstrated by the appearance of many of them amongst the first 1000 sources of the *OED*. In our paper we propose a corpus-based study of their lexical competence to assess their impact on the development of the English vocabulary 1500-1650. This is a pilot study intending to test the use of "sources" in the *OED* for corpus-building, and to combine digital databases and corpus-query systems (*OED*, *EEBO*, SketchEngine) for the diachronic study of lexis. Our study points out a prevalence of church-related vocabulary as a specialised terminology, but it also focuses on other secondary domains such as demonyms and geographical terms.

Keywords: Reformation, corpus-based lexicography, church-related vocabulary, geography.

1. Introduction

As is well known, the early modern period is a key moment in the development of the English lexicon, during which the vocabulary displays the fastest

[1] Both authors are responsible for the overall planning and research for this paper. In particular, Angela Andreani is responsible for sections 2.1 and 3.1 while Daniel Russo for section 2.2 and 3.2. Sections 1 and 4 were written jointly by the two authors.

DOI: 10.25951/4820

growth with a peak observed between 1570 and 1630 (Görlach 1991: 136-137; Barber 1997: 219; Nevalainen 2000: 336; and Durkin 2014: 305-306). Word-formation and increased borrowing both contributed to this growth. The use of written English for most purposes, the expansion of literacy and the spread of printing, combined with the increased mobility to and from England, pushed the creative potential of the language and nurtured a continuous influx of new concepts and foreign words. Since Latin had remained the main language for theology, scholarship and the church for centuries, English had not yet fully developed the vocabulary nor the style of religious debate in a highly dynamic religious context; however, during the early modern period this changed, as the vernacular came to be used "in an increasing range of functions, especially as a language of learning and of religious discourse" (Durkin 2014: 306). As has been remarked, "Nothing reveals the deficiencies of a language more surely than translating into it" (Kay – Allan 2015: 14), which suggests that translators of the Tudor and Stuart era were at the forefront of processes of lexical enrichment. Not only was an immense body of classical Greek, Latin and Hebrew sources turned into English during this period, but the entire vocabulary of the church and religion was discussed, re-codified, and significantly enriched, also through the contact with other vernaculars: "previously 'dogmatic' words like *heresy*, *enormity* and *abuse* became relative and plural in meaning, as their use became dispersed among the disputants" (Hughes 1988: 113). Intending to map the influence of religion onto the history of the English lexicon, in this paper we try out a combination of resources and methods that will enable us to explore the intersection between lexicography, translation and religious writing.

2. Materials and methods

2.1 Materials and sources

Our materials were retrieved starting from the *OED's* top 1000 sources and identifying a group of translators, theologians and controversialists active between 1500 and 1650. The data that can be accessed using the *OED's* sources have already proven valuable for linguistic research; Giles Goodland (2013) has investigated the use of neologisms in early modern literature by focusing on a selection of canonical authors retrieved using the "sources" function of the *OED*, while Julie Coleman (2013) has shown what can be gained from a close analysis of the *OED's* sources combined with an awareness of the limits of this function.

In fact, working with the *OED's* sources opens up a number of methodological questions. One of the main concerns for scholars is the representativeness of the quotations used in the *OED* (Schäfer 1980; Brewer 2010 and 2013; Considine 2009; and Coleman 2013); however, this limit of the *OED's* sources does not prejudice our research, since our starting point is the study of the contribution and legacy of a selected category of writers. Another limit, pointed out by Charlotte Brewer, is that the data searched through the *OED* are not stable since "every quarter, the identical search will produce a different set of results, as the lexicographers upload a new batch of revised entries to the dictionary and remove the corresponding unrevised ones" (2013: 115). This means that there may be discrepancies between our data and the information published on the *OED Online* when the entries involving our source authors are revised. In fact, we may have spotted a couple of such instances working with our data (see §2.2).

Our selection of authors was based on background knowledge and on information we could verify using the *ODNB*. From the list of the *OED's* sources we selected authors whose written output and profession indicates lifelong interaction with Biblical and patristic sources, in the original or in translation. Included are works that cannot be classified as translations proper, and people to whom the professional label of translators cannot be applied. In fact, what constitutes translation, citation or paraphrase in this period is fuzzy, but our assumption is that operating across languages and cultures was standard intellectual practice for our authors. Our sampling includes a combination of established and less canonical figures. In chronological order, our source authors are:

- John Bale (1495-1563), reformed clergyman, bishop of Ossory in Ireland, active evangelical polemicist and author of a commentary of the Book of Revelation;
- John Foxe (1517?-1587), the renowned author of the *Acts and Monuments*, he had a deep knowledge of early Christian historians on which he based sections of his work, was the author of controversial tracts and collaborated with Continental reformers;
- Thomas Cooper (c. 1517-1594), bishop of Winchester, theologian, editor of Thomas Elyot's dictionary and himself an important English-Latin lexicographer;

- John Jewel (1522-1571), bishop of Salisbury, the chief apologist of the Church of England who confuted the authenticity of the Roman Church based on the Patristic sources of the early centuries of Christianity;
- John Daus (c. 1516-1602), chaplain and later schoolmaster and preacher, translator of the works and sermons of prominent European Protestants, and allegedly of Eusebius' *Ecclesiastical Histories*;
- Arthur Golding (1535/6-1606), translator of a series of major works by Calvin, of the Lutheran commentaries on the New Testament from Latin and of numerous works by Continental writers such as Beza, Bullinger, Augustin Marlorat and Philippe Duplessis-Mornay;
- William Fulke (1538-1589), one of the most important controversialists of the Elizabethan age, chaplain and college head who published an extensive confutation of the Rheims translation of the Vulgata in English and engaged in controversy over the translation of the Bible;
- Richard Hooker (1554-1600), clergyman, deputy professor of Hebrew at Oxford, and the most prominent theologian of the period, author of *Of the Laws of Ecclesiastical Polity*;
- James Bell (d. 1606?), translator from Latin of religious writings by John Foxe, Martin Luther, and Walter Haddon;
- Thomas Newton (d. 1607), clergyman and translator, the most eclectic in our sources, he translated and published on a wide range of subjects, mainly secular such as translations of Cicero and Seneca, was also author of *An Herbal for the Bible*;
- Thomas Tymme (d. 1620), a clergyman who published translations of theological works and devotional writings from Latin, French, alongside his own devotional writings;
- William Sclater (c. 1575-1627), clergyman, author of several sermons and of a treatise on justification, best known for his Expositions of the Thessalonians;
- Thomas Taylor (1576-1632), clergyman and a very prolific writer, author of several sermons, religious treatises, and a commentary of Paul's epistle to Titus.

2.2 Method

The study is based on the concepts, frameworks and methods of corpus-based terminology (Cabré 1998; Gamper – Stock 1998), corpus-based analysis of language variation and use (Biber 2009), specialised discourse (Gotti 2005) and specialised translation (Gotti – Šarčević 2006). The methodological framework of this research project rests upon the extraction and analysis of terminology from a lexical source, i.e. the *OED Online*. Ahmad – Rogers (2001: 584) define automatic term extraction as "the processing of texts using computer programs in order to identify strings that are potential terms"; the most valuable result of term extraction is thus the lexical material that can be used to create terminology databases through a process of examination, testing and validation before items are inserted into lexical resources such as dictionaries. If structured collections of texts are an extremely important source of data in the study of terminology for indexing purposes, one may question the validity of extracting terms that were in turn extracted and processed by the compilers of a dictionary. Rather than showing the contribution of one specific author, the aim of this project is to show the lexical impact of a profile of scholars in the early modern period, especially in unexpected lexical areas; thus, even though this study might be affected by the same possible biases in selection criteria of the *OED's* compilers (Coleman 2013), we believe that working with big data (Weikum *et al.* 2012) and fuzzy sets (Ma 2011) can compensate for this issue.

Extensive research has been conducted on the methodology for lexical extraction (see Pantel – Lin 2001; Jang *et al.* 2021), which according to Mei *et al.* (2016) can be assigned to three macro approaches: rule-based methods, statistical methods and hybrid methods. In the rule-based methods, words are extracted from a lexical resource (a text, a corpus of texts or a dictionary) based on predetermined criteria, which can be linguistic in nature (e.g. morphological categories), but also textual (author, topic, date, etc.). This method is particularly suitable for extracting new or unindexed words (Isozaki 2001; Stanković *et al.* 2016). Statistical methods are based upon statistical linguistic features and usually pair up with machine learning algorithms to extract words in vast corpora; this method is particularly effective when studying collocations (see Pecina 2010) or linguistic patterns and semantic shifts (see Boukhaled *et al.* 2019). Hybrid methods are the combination of rule-based methods and statistical methods and are mainly employed in text mining in language-specific domains (see Hadni *et al.* 2014). This paper is based on a rule-based approach, the rule being *OED* entries listed as first citations assigned to a pre-established list of authors (see §2.1). For most rule-based methods, the definition of rules may be a difficult task resulting in poor systemic flexibility, but as the selection rules employed for this paper are domain-specific and extralinguistic, this issue does not arise.

Paraphrasing Wright – Budin (2001: 726), text corpora are a valuable source of evidence when studying the variation of occurrence and use of a language for specific/special purposes (LSP) and its terminology in three main aspects: across specialistic domains, between levels of communication, and diachronic change in relation to competing morphological forms, spellings, and terms. Although our lexical extraction pursues the same purposes, we prefer referring to our collection of lemmas as a *database* and to the proper collection of texts of the *OED's* compilers and to the reference collection of early modern English books as a *corpus*. For decades, the notion of corpus has been understood in linguistics as an electronic corpus, which is stored, processed and analysed automatically or semi-automatically by specialised software systems (Baker 2006: 25-26). While commenting on the relationship between lexicography and translation theory and practice in Tognini-Bonelli (1996), Hanks focuses on the distinctions between corpus-based and corpus-driven lexical research: the aim of corpus-based is to force the lexical evidence of a corpus to fit into preconceived theories through the use of "judiciously selected examples", whereas corpus-driven studies attempt to approach data "with an open mind and to formulate hypotheses and indeed, if necessary, a whole theoretical position on the basis of the evidence found" (2012: 417). On the other hand, Xiao (2008) maintains that this sharp distinction found in the literature between the corpus-based approach and the corpus-driven approach is largely overstated. We support this less polarising view at least for the purposes of this project: as will be described in detail below, the extraction phase of this pilot study is completely corpus-driven; however, the analytical phase must be corpus-based as the texts and the lemmas are examined diachronically and belong to a period wherein terminological approaches, writing practices and semantic prosodies varied considerably; a more manual analysis is thus fundamental to establish synchronic and diachronic connections that would otherwise be overlooked.

The main methodological aspect involved in this study is the extraction of lemmas from the *Oxford English Dictionary Online*. The assessment of any term-extraction method must comport with an evaluation of the corpus that is being analysed, not simply as in traditional terminology management in relation to the authority of the authors of texts, but also with regard to the structure and processing of the corpus (Ahmad – Rogers 2001: 585). The automatic extraction of lexical items is one of the most significant problems in Natural Language Processing (NLP): normally in corpus-based studies the aim of word extraction is to isolate sets of terms and expressions with a certain meaning in a collection of text strings (this only partially applies to the purposes of this paper, as will be discussed below). Applications of computer-aided term extraction include information retrieval, lexicography, parsing, computer-assisted and machine translation, and lexical databases. In an effort to pursue this last application, this paper sets out to build and examine a lexical database of first occurrences of lemmas extracted from the sources listed in §2.1 from the *OED Online*. Instead of an evaluation system relying mostly on human assessments of the quality of extracted terms, we intend to combine automatic extraction and human analysis. The lemmas listed as first occurrences of the authors described in §2.1 in the *OED Online* (section Sources>[author's name]>first entry)[2] are extracted through a Python script (Hammond 2020) and entered into a spreadsheet database, which stored the following data: author, lemma, definition, work-title, date. The script was executed twice in order to verify whether the data were consistent over time – on 4 November 2019 and on 5 February 2021 – and it proved that the great majority of first citations were not amended during the period concerned, only a very small number of first citations had been re-assigned to another (mainly coeval) source during the months from the first extraction to the second, e.g. *Christianlike* was assigned to Newton (1574) in 2019 and to Taverner (1540) in 2021 and *Bohemian* (sense b), formerly Fulke (1579), is now Golding (1562). The second phase involves the classification of the dictionary entries in relation to their semantic field. This process has produced a list of 1,919 lemmas that are indexed with the following information: author, definition, title of first occurrence, date of first occurrence. Although a certain number of entries show some level of classification in the definition section (e.g. "anatomy"), most lemmas do not; therefore, this phase required manual processing, which was time-consuming and, in a few cases, implied terms being classified in more than one category. We have identified several prevailing semantic domains in the extracted database; therefore, in this pilot study we will focus on one expected domain in the corpus, i.e. religious terminology, and on one unexpected domain, i.e. geographical terms. In the third phase, we relied upon another digitised corpus – the Early English Books Online (*EEBO*); through this corpus, which can be efficiently browsed on the online corpus linguistics platform Sketch Engine (Kilgarriff *et al.* 2004 and 2014), we manually evaluated the lexical and semantic aspects (especially in the form of concordances) of the terms in our database in order to obtain a comparative evaluation of term usage in competing forms, spelling variation and, possibly, dating with the extended corpus of early modern English sources.

[2] https://www.oed.com/sources.

3. Results and discussion

This section is divided into two subsections, one for each lexical macrodomain discussed in this paper. The main aspects tackled in our approach are occurrences, spelling, morphology, competing expressions, semantics and etymology. Several intersections are discerned in these seemingly unrelated fields. The years indicated in brackets in the citations below are those reported in the *OED* and extracted into our database; *EEBO* references are indicated with their TCPIDs (Text Creation Partnership ID), which unambiguously identify the source.

3.1 Religion and church-related vocabulary

Church related vocabulary amounts to 177 lemmas ranging from words connected with the writing and the study of the Bible, to words to indicate behaviours against the church, members of the clergy, God, or the sacraments. To make sense of this diversity, the lemmas have been organised into semantic fields and categories, drawing from the classes and senses used in the *OED Historical Thesaurus* (*HTOED*). Three macro-categories have been identified: "faith", the "supernatural", and "other". "Supernatural" defines words and attributes for God and deities (i.e. *petty goddess*, *theandric* and *unitrine*[3]); "other" includes a variety of words from various semantic fields, which have developed (or preserved) senses connected with religion and the church (e.g. *church story*, *disvesture*, *ministership*); and "faith", the first category with 144 lemmas, includes the fields in Table 1 below, arranged from the most to the least numerous:

Table 1. Church-related vocabulary > "Faith": fields and nr. of lemmas

Field	nr. of lemmas	Field	nr. of lemmas
sects	41	canon law	2
church government	33	creed	2
sacrament, communion	9	architecture	1
paganism	8	prayer	1
liturgy, ritual	8	error	1

[3] For the purposes of this paper, words extracted from our database are marked in italics, whereas glosses, definitions, translations, mentioned words and phrases, etc. appear between double quotation marks.

sacrament, ordination	5	Catholicity	1
scripture	5	atheism	1
consecration	4	heresy	1
benefices	4	offence	1
apostasy	3	orthodoxy	1
sacrifice	3	religion	1
canonization	3	sectarianism	1
sacrilege	2	spirituality	1

The data photograph a very rich and composite situation. In what follows, the discussion will be limited to selected examples illustrative of lexical enrichment in particular fields, with attention to morphological experimentation and semantic shifts.

The number of lemmas that indicate religious sects is staggering. In our database we have a variety of new entries and derived forms. A number of lemmas are based on aspects of discipline or of doctrine, such as *flagellant* and *anabaptistry*, and several are derived from the names of their founders, such as *Arianism*, but also *Christianlike*, *Calvinist* and *Mahometical*. In two cases, provenance defines particular sects: *Saxonian* and *Bohemian* (see §3.2 below).

Sometimes we can clearly detect the influence of patristic sources. The noun *Donatian* (Sclater 1627) is a Latin loan whose entrance into English was mediated by the work of Jerome and Augustine. This variant had limited use in the early modern period (12 hits in *EEBO*) and is now obsolete, while "Donatist", much more frequent in the *EEBO* corpus (497 hits), is in current use. *Donatian* may be a zero derivation, since the adjective is attested four decades earlier in the *EEBO* corpus: "as S. Augustine sommetime saide to the Donatian Heretiques" (A04468). *Marcosian* (Fulke 1580) is derived from Greek and like the names of several other sects it entered into English through the popular early Christian work on heresiology written by the bishop of Lyon Irenaeus (*c.* 130-202), *Adversus Haereses*: "Transubstantiation of the wine into blood in Marcus and the Marcosians *Irenaeus lib. 1 cap. 9.*" (A01325, original italic). Another channel was the compendium by the bishop of Salamis Epiphanius (d. 403) known as the *Panarion*, or *Adversus Haereses*: "Likewise the Marcosians when they baptized, vsed to speake certaine Hebrue wordes, […] *Epiph. lib. 1. Tom. 3. haer. 34.*" (A01335, original italic).

The close contacts established by English reformers with Continental communities favoured influences across the vernacular languages.

A number of terms of classical origin may have been modelled on coeval forms in French, German or Italian; this was the case for the term *Confessionist* (Fulke, 1570, but the earliest occurrence in *EEBO* is 1565 in A04474), from French *confessioniste*, used as a synonym for "Lutheran", although much less frequently (23 vs 6,635 hits). Another example may be the term *Calvinist* (Fulke 1579), for which we find the equivalent *calviniste* in French.

Examples of coinages from internal derivation processes are *Anabaptistry* (Foxe 1570), from "Anabaptist", *Lutheranism* (Daus 1560) from "Lutheran", *Puritant* (Fulke 1580) from "Puritan", and possibly *Calvinist* (Fulke 1579), which may have been modelled on the slightly earlier "Calvinism" or derived directly from the name of John Calvin (see *EEBO* A20661 for earlier occurrences dating to 1564, e.g. "how happeneth it that the Caluinistes and the Lutheranes agre not").

Finally, the data illustrate to what extent the separation from the Church of Rome triggered the lexical inventiveness of polemicists. "Popery" became a derogatory catchword for Roman Catholicism, which was framed as a false and idolatrous religion in sermons, pamphlets and treatises. The first lemma in our database is *papistry* (Bale 1543) derived from "papist" (OED s.v. "papist, A. n. 1"), and with its 901 hits in the *EEBO* corpus the most frequent keyword for anti-Catholic slander coming from our sources:

(1) not onely defending the vngodly worship, papistry, and false religion. (*EEBO* A04696)

(2) euen so they that are droonke with the hereticall doctrine of Papistry. (*EEBO* A01327)

(3) The religion of papistry being a Catholick Apostasie from God. (*EEBO* A20740)

The oldest and most frequent alternative by far is "popery" (the spellings popery + poperie retrieve 26,073 hits, with the earliest attestation dating to 1528), not present in our database. A borrowing from Latin, *papism* (Bale 1550) is another relatively frequent term of polemical slander (328 hits), while the adjectives *popan* (Bell 1580) and *papane* (Bell 1581) are variants for the much more frequent and established "popish" and "papal" (respectively 23,410 and 6,041 hits in *EEBO*); in particular, *popan* might be a nonce form by Bell, but a search of *papane* (overall 15 hits) retrieves earlier attestations of the term in the sense of "Pope's dominion" (A17662) and as a synonym for "papal" (A01130).

Morphology reveals processes of selection and acceptance in the history of several words: in our database *Lollery* (Bale 1547, 4 hits including the spelling "lollerie") occurs as a variant for "lollardry" (4 hits) and "lollardy" (39 hits including the rarer spellings "lollardye" and "lollardie"). In the *EEBO* corpus "Calvinian" appears alongside the variant from our database *Calvinist* (Fulke 1579) although with less freqency (520 vs 1767 hits respectively). With 56 hits *Wycliffian* (Foxe 1570) supersedes the alternative variants "Wycliffist" (18 hits) and "Wycliffite" (21 hits); it may be noted that all appear with that they spelling "Wick-" in *EEBO*.

In order to place the terminology denoting religious sects in the broader context, we have turned to the *HTOED*, which reveals that these terms entered the English vocabulary from different channels and into stages. A consistent portion entered through the translation of the medieval collections of saints' lives, while the 16th century additions may be explained in part as the effect of the recovery of patristic sources and their translation into English, and in part with the need to make sense of an increasingly fragmentary religious situation through lexicalisation. Another semantic field that emerges from our database is in fact that of "sectarianism" (e.g. *interimist*, Daus 1560). If we expand the search for related terminology in the *HTOED* we see that cognate words (e.g. "sectary", "sectator", "sectuary", "sectist") and semantically related forms such as "separatist", "conventicler", and variants, are all additions dating between the 1550s and 1600, signalling the particular development of this area of the English lexicon during the period under review.

Our database highlights another area of special significance in the lexical repertoire of religious authors and translators: Eucharistic terminology. From Latin we have the verb *transcorporated* (Foxe 1570), seemingly a nonce usage proposed as an alternative to the older and more common "transubstantiated" (511 hits in *EEBO*, earliest attestation 1549). The verb *inaccidentated* (Fulke 1579) in our database appears to be another nonce usage. Neologisms of this kind seem to be a distinctive feature of controversial literature; compare the term "iniesuated" (*EEBO* A02617) and further interesting coinages by William Fulke present in our database:

(4) but he [i.e. Christ] is not to be worshipped in bread & wine, or in *the* accidents of bread & wine, because he is neither impanated, nor inuinated, nor inaccidentated, that is, not ioyned to any of them in a personall vnion. (Fulke 1579).

Inaccidentated was derived by affixation from the word "accident", perhaps after the model of the loan *impanated*, which occurs alongside *invinated*, introduced by Fulke in 1579, and likely derived from an earlier form "invinate" already attested in 1550 (*EEBO* A19571).

The occurrence of the pair *consubstantiation/consubstantiate* in our corpus (Hooker 1597) reflects the development of the Eucharistic debate. These words were in fact specialist controversial terminology, in that they helped define and identify different theological positions, as the citation from Hooker's *Of the laws of ecclesiasticall politie* makes clear:

(5) They […] are driuen either to Consubstantiate and incorporate Christ with elements sacramental, or to Transubstantiate & change their substance into his. (Hooker 1597)

and further:

(6) So that they all three do plead Gods Omnipotency: Sacramentaries, to that Alteration, which the rest confess he accomplisheth; the Patrons of Transubstantiation, over and besides that, to the change of one substance into another; the Followers of Consubstantiation, to the kneading of both Substances, as it were, into one lump. (*EEBO* A44334)

These examples enable us to appreciate how morphological experimentation could convey key religious meanings: neologisms with prefix *in-* and denominal suffix *-ate* could be used humorously and/or to convey polemical and derogatory overtones (example 4); the prefixes *con-* and *trans-* could encode specific doctrinal positions regarding the understanding of the body of Christ in the communion (example 6). As a process of word formation, therefore, derivation is not only particularly productive but also of special significance in the lexis of religion, as beliefs, groupings and outlooks became lexicalised.

In the iconoclastic setting of Tudor and Stuart England, words referring to images acquired negative connotations too. An entire vocabulary derived from originally neutral terms such as "image" and "idol" became bywords for paganism, heresy and a false Christianity. *Idolatrous*, whose first evidence is provisionally found in Bale in the *OED* (1550, see s.v. "idolatrous, adj."), is in fact an older presence in the English vocabulary: "a supersticious and idolatrous kynde of worshippyng" (1542, *EEBO* A06710). This variant is the one that has become established in English, and with 9,606 hits in *EEBO* it proves to be already well-attested in our period. Both apparently introduced by Bale, the adjectives *idolous* (Bale 1546), and *mammetrous* (Bale 1546) are considerably less frequent with 5 and 1 hits respectively. *Mammetrous* is derived from the 14th century loan "mammetry", which indicated idolatry and non-Christian practices. The term is in fact a loan from Anglo-Norman *maumeterie*, a reduced form of *mahumetterie*, ultimately derived from the name of the prophet Muhammad (*OED* s.v. "mammetry, n."). By the time they entered English, *mammetrous* had evidently lost all connections with Islam, so that in our database we find the new entries *Mahometical* (Daus 1561), borrowed from French and Latin, and *Mussulman* (Foxe 1570) borrowed from Persian, Arabic or Turkish (*OED* s.v. "Mussulman, n. and adj."). The new loans are not associated with idolatry, as may be expected, but are nonetheless connotated as blasphemous practices by our sources: the phrase "Mahometicall corruption" appears in a translation of Bullinger's sermons (Daus 1561), and *Musulman* as a term for a "Turkishe priest" (Foxe 1570, on *Turkish* see 3.2 below). With 154 hits in the alternative spellings *Mus*(*s*)*ulman*(*s*), this term superseded *Ma*(*c*)*hometical*(*l*) with its 51 hits. One final coinage in this field, the compound *image-worshipping* (Bale 1544), appears to have been used very limitedly (10 hits) in comparison with the older and well-established "idolatry", a borrowing via French (over 40,000 hits).

Words that have been grouped under the field "church government" display processes of pejoration, especially those related to the field of monasticism, such as *abbey-like* (Foxe 1570) and *cloistered* (Bell 1581), which show that monastic lodgings were framed as places of corruption: "Shewing, The Canterburian Cathedrall to bee in an abbey-like, Corrupt, and rotten condition" (*EEBO* A35353); "these Cloistered Friers, who now grown to the height of their sinnes" (*EEBO* A12738). The word *greasling* (Golding 1583), a derogatory term for Catholic priests derived from "greasing", used contemptuously to refer to the practice of "anointing" in religious ceremonies of the Roman Catholic church:

(7) their popish greasing which they vse only when a man is desperatly sicke. (*EEBO* A01325)

Another interesting lexeme used in our database to denote priests of the Roman Catholic Church is the compound formed within English *mass-monger* (Bale 1551), denoting a "dealer" or a "trafficker" in masses:

(8) For our Massemongers haue Masses in store for all kynde of thinges good or badde. (*EEBO* A06652)

Compared with its current meaning, the term *seminarist* (Fulke 1583) in our data has markedly negative connotations, clearly due to its association with Roman Catholicism:

(9) than all the popish Seminaries, and Seminarists, shall be able to hinder it, iangle of grosse & false translations. (Fulke 1583)

This term and the more frequent compound "seminary priests" are often paired with "Jesuit", when not used as synonyms:

(10) These Seminarists Jesuits, and other Priests. (*EEBO* A20820)

In the late 16th century, they represented in fact the quintessential seminary priests, trained on the Continent, especially at the English college of Douai, which, since 1574, had been the fulcrum of the reorganisation of militant English Catholicism:

(11) the flocking of so many Iesuits and Seminaristes, as so many trompets and bellowes of sedition into England. (Fulke 1583)

In its current sense of "member of a seminar" (OED s.v. "seminarist, n."), the term has undergone secularisation. Other examples of secularisations concern the words *customariness* (Cooper 1608), originally denoting "perfunctory worship", but whose extended use is already attested in the 17th century, and *renouncer* (Bale 1547), denoting especially renouncers of God, the Truth, or the Church and often paired with "abiurer" and "apostate" in the *EEBO* corpus. A common adjective in Present Day English, *ritual* is another term from our database (Foxe 1570) that may be said to have undergone secularisation, as it originally referred to the performance of rites, often intended as empty ceremonies:

(12) Of these solemnities & feastes we reade that they belonged & were inioined to the Iewes vnder the law, were meerly ceremonial & ritual, […] neither are to be reteined in the church or ministerie of CHRIST. (*EEBO* A05025)

One final example worthy of attention is *superintendent*, another of several terms attributed to Bale. According to the *OED*, this is a loan from post-classical Latin after the ancient Greek *ἐπίσκοπος*, found in Jerome to indicate a "superintendens bishop". In Continental churches and in the reformed Church of Scotland it denotes a chief presiding minister (still in use), and an official appointed to ordain ministers and to oversee a territory (obsolete), but the sense "superintendens bishop" is specific to the English context, and it was used by both reformers and Catholics, with opposite connotations, to indicate the bishops of the Church of England. The term in English was apparently modelled after German *Superintendent*, and its adoption shows both the influences across vernacular languages and the use of early Christian texts as sources for a vocabulary to describe the reformed Church government. A series of examples retrieved from *EEBO* illustrate the gradual establishment of the term "superintendent" in our period, through definition, synonymy and explanation:

(13) Episcopus is as moche to saye as a superintendent or an ouersear, whose offyce was in the prymatyue Churche purelye to instructe the multitude in the wayes of God. (Bale 1544)

(14) And the word (superintendent) being a very latin word made English by vse / should in tyme haue taught the peple by the very etymologie & and proper signification. (*EEBO* A10777)

The final example is particularly cogent coming from an author clearly of Catholic leanings attacking the Reformation as (also) terminological subversion:

(15) They had throwen doune altars, ouerthrowen Churches, denyed all outward Priesthod, changed Bishops into superintendents, Priests into ministers, altars into tables, the chaste clergy into the vnlauful mariage of votaries […] (*EEBO* A11445)

3.2 Demonyms and geography

The total number of lemmas concerning geographical entities is fifty; amongst these lexical items we have further identified the following subcategories: toponyms, demonyms, geographical entities, geo-political institutions, expressions deriving from geographical references.

Oddly enough, proper place names are not significantly represented in the database, with a total of three entries, all of them being related to classical Greco-Roman heritage. In the preface of the English translation of Levinus Lemnius's *De habitu et constitutione corporis* (1561) by Newton (1576) we can read the placename *Camaryne* (today's Camarina in Sicily, Italy) in the obscure idiom "wade into the very Gulphe & Camaryne of mannes apparaunt wilfulnesse". In order to decipher this expression, it is necessary to read Strabo's account of the marsh of Camarina (in Jones 1978 [1917]: 59-82): before the Carthaginians destroyed Camarina in 405-401 BCE, its inhabitants were plagued by malaria caused by a nearby marsh; once they dried it, the disease stopped spreading; however, there was no longer anything stopping Hannibal's army from razing the city. In this sense, *Camaryne* becomes the metonym of "marsh", and the term is thus defined by the *OED* as "a fetid marsh or swamp". The next place name that can be identified in the database is *Sarum*, which first occurs in Foxe (1570) in various collocates: "dioces(se) of Sarum", "Bishop of Sarum", "Chancellor of Sarum", etc. both in English and in Latin. *Sarum* is a latinised form of "Sar" a medieval abbreviation of Salisbury (Mills 2003); both in Foxe and in other contemporary works found in *EEBO* (e.g. A07139, A16292, A05547), *Sarum* does not only refer to a geographical entity but more specifically to the so-called Use of Sarum (or Sarum Use), i.e. the Latin liturgical rite developed at Salisbury Cathedral from the late 11th century until the English Reformation (Cheung Salisbury 2009). Even more prominently, the third toponym extracted in the database is highly symbolic, *Sodom*, which appears in Bale (1550); in this text, the biblical reference, which is also spelt as "Sodome" and "Sodoma", is mainly used to portray Rome as the place of Papal corruption (16):

(16) why so tyrannouslye bynde ye them, to that fylthye Sodome, withoute redempcyon? (Bale 1550)

In *EEBO* the spelling "Sodom" outnumbers (7,092 hits) both "Sodome" (3,048 hits) and "Sodoma" (211 hits), which confirms today's spelling.

What stands out when browsing the list of geographical lemmas is certainly the peculiar prevalence of demonyms and adjectival phrases related to geographical entities, often with competing variations. Considering the religious background of the time and the main influences in the Protestant reform, there is a high occurrence of first citations in lemmas related to the German-speaking areas of Europe. First and foremost, the word *Dutchland* in Bale (1547), which in *EEBO* seems to be a calque from German *Deutschland* and a less common alternative (99 hits) to "Germany" (31,024 hits), "Germanie" (5,216 hits) and the French inspired "Almaine" (1,048 hits) and “Alemaine” (20 hits). The fact that these place names were perceived as synonyms is clear in “Germany, is a country called of some Dutchland, of some Almaine” (*EEBO* A05237). However, the confusion between this old and the current meaning of “Dutch” is also apparent in the corpus, and the authors of our database also provide first occurrences of two adjectival phrases that are used to distinguish between Germans and the Dutch: *High Dutch*, i.e. High German or *Hochdeutch*, which is found in Daus (1560), and *Low Dutch*, i.e. Low German or *Niederdeutch,* which first appears in Newton (1576). The distinction between these varieties of Germanic languages is widely understood in the scholars of the time, as can be observed in (17):

(17) Although I bee well acquainted with the high and low Dutch tongue, yet I must confesse that in this ancient Frison language I vnderstand nothing. (*EEBO* A68345)

Other German-based competing demonyms in the database are *Saxonian* in Hooker (1599) and *Saxonish* in Bale (1549); the former carries a geographical yet religious connotation in Hooker (18); for this reason, the *OED* assigns the definition “a Protestant of Saxony” to this entry. Moreover, these competing variations have low scores in *EEBO*: *Saxonian* 11 hits (which *OED* lists as an obsolete form of *Saxon*), *Saxonish* 10 hits (marked as archaic in the *OED*), whereas *Saxon* has 22,896 hits in adjectival phrases.

(18) the French Protestants took Arms against their King, […], the Belgick, the Helvetian, the Bohemian, the Saxonian, the Swevian, the English, as consenting for Obedience to their Soveraigns. (*EEBO* A27046)

Similarly, in his English translation of Sleidane’s *Commentaries* (1560), Daus uses both *Suevical* and *Swevical* to identify Swabian Protestants, but this appears to be his own coinage, as there is no other evidence of these adjectives in works other than the *Commentaries*. Furthermore, in the same translation Daus refers to Slavic peoples as *Slavonish*, which has only two other occurrences in *EEBO*, mainly in relation to the Slavic peoples settled in the Balkans. Our database also contains another term denoting a Slavic people – *Bohemian* – in Golding (1562), as can be seen in (19); once again, this term designates a geography-based religious entity, as these “Bohemians” are Bohemian Protestants, or Hussites. *EEBO* lists 1,119 occurrences of *Bohemian* alongside the competing variation “Bohemish” (5 hits).

(19) Thus we are hable to allege Luther, Melancthon, Bucer, and that learned Bohemian, for the indifferencie of the Communion to be ministred either vnder one kinde or bothe. (Golding 1562)

The second large area covered by the first occurrences in the geographical section of our database includes the lands and countries bordering the Mediterranean Sea. A small number of first citations concern places in Italy, with obvious references to the conflict with Roman Catholic clergy, e.g. in Foxe (1563) *Etruscan* is found in the phrase "Etruscan tyrant", which needs to be contextualised. Foxe portrays Bishop Bonner as the perpetrator of the most vicious cruelties and injustices against English Protestants under the Catholic government of Mary I of England; his victims included Thomas Tomkins, whose hand was burned following the bishop's orders. Tomkins' faith was tested by Bonner; likewise, Scaevola's valour was tested by the "Etruscan tyrant" Porsenna: Foxe employs this comparison as a means to invest Tomkins with a heroism comparable to that of a legendary champion. The adjective *Italish* stands out in the database as a first citation in Bale (1544); however, in *EEBO* this appears in two collocates that can be traced back solely to Bale (1548) – "Italish warre" and "Italish préest" (A68202) – thus we can conclude this form is likely to be his own coinage as an alternative to "Italian", which occurs extensively in the same pages (e.g. "Italian prouerb"); however, in this text "Italian" prevails mainly as a noun, e.g. "in the yéere of Christ 1368: which yéere the Italians count 1367" and "the ambassador of France was also present with another stranger an Italian". This distinction is not confirmed in *EEBO*, where "Italian" occurs as both a noun and adjective as in contemporary English. Two competing adjective forms are present for Adriatic Sea: *Adriatical*(*l*) in Cooper (1549) and *Adrian* in Newton (1575); as might be expected, in both cases these adjectives collocate only with the noun "sea". *EEBO* shows that these adjectives were indeed competing variants: *Adriatic* has 38 occurrences, *Adriatical*(*l*) 20 occurrences, and *Adrian* (sea) 15 occurrences; in modern usage, *OED* marks "Adriatical" as obsolete, "Adrian (sea)" as poetic and rare. A similar consideration can be made in relation to the word *Turcian* that appears in Foxe (1570): whilst *Turcian* seems to be a nonce word in *EEBO*, two other forms are in competition – "Turkish" (4,629 hits) and "Turkic" (3,165 hits), which in contemporary English ended up conveying different meanings ("relating to Turkey" and "related to the Turkic language family" respectively). The *OED* also lists a very peculiar usage of "to turkish" as a verb meaning "to transform, especially for the worse; to pervert; to turn into something different" from (20):

(20) sayeth how the turkyshed seede is sowen abroade in England, and in Germany, signifying the doctrine that is contrary to the byshop of Rome. (Daus 1560)

Finally, there are a few other first citations belonging to the Mediterranean area: *Mozarabical* in Newton (1575); *Costantinopolitan* in Fulke (1577); *Ephesine* in Fulke (1555); and *Hierosolymitan* in Bale (1538), the last three being originally geographical terms modelled after Romance adjectives and used in these writings as religious references to Christian denominations and ecumenical councils. In *EEBO*, there is only one occurrence of *Mozarabical* by Newton (A19712) alongside 12 occurrences of the competing form "Mozarabic(k)", which mostly collocate with "liturgy", "use" and "office" to identify a liturgical rite of the Latin Church once used generally in the Iberian Peninsula; there are only two occurrences of *Constantinopolitan*(*e*) as purely geographical references; there are 520 occurrences of *Ephesine* and 355 occurrences of the competing form "Ephesian" (which would later become the primary adjective referring to Ephesus); there are 31 occurrences of *Hierosolymitan* and one occurrence of "Jerusalemite" (which is today's most common adjective relating to the city of Jerusalem). It needs to be noted that these words are still used today although they are in some cases marked as dated, but their semantic value have shifted from mere geographical to mostly historical and religious.

A special mention needs to be made for the first occurrences of terms related to Graeco-Roman geography in Golding's translations of Caesar's *Commentaries* (1563), Ovid's *Metamorphoses* (1565), and Pomponius Mela's *Geography* (1583). These expressions include *Parnassian*, from the Greek mountain Parnassus; *Pylian*, the inhabitants of the ancient Greek town of Pylos; *Pythian*, the demonym of Delphi (whose ancient name was Pytho), whose root allegedly derives from the word "python" in (21); "Salentine", the demonym of the ancient tribe of Messapians, also known as Sallentini in ancient Rome (22).

(21) Python […] Which of the serpent that he slue of Pythians bare the name (Golding 1583)

(22) Spartanes buylt, and Cybaris, and Neaeth salentine, And Thurine bay, and Emese, and éeke the pastures fyne Of Calabrye (Golding 1583)

In this translation of the *Geography* we can also find mentions of the *Seres* (390 occurrences in *EEBO*) along with the correlate adjective *Seric*, which are respectively a loanword and a calque of the Greek and Latin *Seres/sericus* (ultimately from the word "silk" in various Eastern Asian languages, wherefrom the current English word "seric" derives) to identify the Chinese, as in (23):

(23) We vnderstand that the first men in Asia Eastward, are the Indians, Seres, and Scithians. The Seres inhabite almost the middle part of the East, the Indians and Scithians, the two vttermost partes: both peoples extending farre and wide, and not onelie toward the East Occean. (Golding 1583)

In addition to this, our database includes the first mention of the demonym *Asian*(*e*) as a noun (1,225 occurrences in *EEBO*): it is found in Bale (1548) in (24):

(24) These were of all nacions of the earth, of al peoples of the world, and of all languages vnder heauen, Gréekes, Latines, Hebrues, Caldeans, Parthyans, Medes, Elamites, Capadocians, Asianes, Phrigian, Egiptianes, Arabianes, Syrians, Africanes and Indians. (Bale 1548)

Our database is also populated with a relatively small number of first citations of foreign local institutions, most of which are borrowings or calques from contemporary non-classical languages. The most remarkable case is the triplet *Sorbonne*, *Sorbonist* and *Sorbonical*(*l*) – the first and the second occur in Daus (1560), the third in Bale (1543) – which highlights how deeply the Sorbonne became involved as a reputable institution with the intellectual struggle between Catholics and Protestants in the 16th and 17th centuries (Conway 2009). Other expressions in this section include *Archduchy*, more specifically the "Archduchy of Austrich" (Foxe 1563) which was possibly a calque of French from Latin (in *EEBO* this term occurs exclusively in the collocation "Archduchy of Austrich", "Archduchy of Austria" and "Archduchy of Insbruck"); *burgrave*, a calque from German *Burggraf* (a military governor of a German town of castle in the Middle Ages), which first occurs in Bale (1551) and is found in *EEBO* in 64 concordances in the collocates [burgrave] + [of] + [German city]; *calfam*, probably a corrupted version of "caliph" is found in Bale (1550); *vaivode*, a borrowing from Slavic *воевода/vojvoda* ('army leader' or 'duke'), appears in Daus (1560) and *EEBO*'s concordances show a prevalence for the collocation [vaivode] + [of] + [Valachia/Transilvania] (with one curious exception "Vaivode of Athens"); *vergobret*, a magistrate in ancient Gaul, which appears in Golding (1563); *piazza*, a borrowing from Italian, in Foxe (1583).

Finally, on a more trivial note, our database comprises the first occurrences of the words *bugger* in Daus (1560) and *buggerage* in Bale (1548). Although these words have nowadays lost any spatial reference, these terms originally have a geographical connotation, more specifically Bulgaria, wherefrom the Bogomil heretics were thought to have originated and spread around the 11th century; abominable rituals were imputed to Bogomils and this association is still rooted in today's use of the word. Considering the time frame, the very nature of the works in the corpus, and the profile of the authors under consideration, we support the idea that Daus and Bale cannot have been completely oblivious to this connection.

4. Conclusion

The research hypothesis of this study is that translators, theologians and controversialists active between 1500 and 1650 were leaders in processes of lexical enrichment. This is supported by the data stored in our database of first citations, as shown in the list of first occurrences discussed in this paper. Not only were these scholars innovative in their own field of expertise, but they influenced terminology in a variety of domains, as the examples in the realm of geography showed. They were even confident enough in their abilities to control the morphological aspects of lexis that produced a variety of (co-existing) possibilities, as can be seen in the analysis of the vocabulary of the Eucharist and of controversial neologisms as well as in the adaptation of the loans for religious sects and in the analysis of demonyms that emphasised how suffixes denoting entities belonging to countries, nations and regions from different linguistic sources – i.e. *-ish* (Germanic), *-*(*i*)*an* (Norman French), *-ic*(*al*) (Latin) – used to be employed interchangeably, to some extent at least. The set of terms considered in this study shows a clear prevalence of derivational strategies (60%) especially in the religious vocabulary with a number of instances of compounding (3%) adaptations (10%) and borrowings (29%), these two being predominant in the geographical terminology. Moreover, our data highlight a dense network of influences from the classical languages into English and between vernacular languages, through the sustained contacts of English and Continental reformers and translators: 39% of the words come from Latin or Greek, 17% from Romance languages (mainly French but also Italian and Spanish), 3% from German and 8% from other languages. This emphasises that the vocabulary of the church, of religion and of the peoples and nations of the world was discussed, re-codified, and significantly enriched throughout this period, during which we see provenance, ethnicity and belief as overlapping notions and a source for terminological creativity, as well as confusion. From today's perspective, we can observe that 42% of the words are now marked as obsolete and/or rare, 10% as historical and/or archaic and 1% as poetic, which can be explained by the circumstantial nature of the religious terminology regarding the debates and controversies of the time and the historical distance between early modern Britain and present-day English-speaking countries. The words still in common use are mainly those that refer to entities that have undergone little or no change in identity (e.g. *Lutheranism*, *Calvinist*, *Sorbonne*, *piazza*), or those that have gained ground amongst competing variants (e.g. *Asian, Sodom*), or those that have been reframed as references that are still significant on historical grounds (e.g. *Etruscan*, *Bohemian*).

As mentioned at the onset, this is a pilot study devised to test the validity of our methodological approach. The next steps in our research will be to expand and update the sources of our database by refining the selection criteria and to investigate further semantic domains emerging from the data.

REFERENCES

Sources

Early English Books Online (*EEBO*)
https://proquest.libguides.com/eebopqp, accessed February 2021

Historical Thesaurus of the Oxford English Dictionary Online (*HTOED*)
https://www.oed.com/thesaurus, accessed February 2021

Oxford Dictionary of National Biography (*ODNB*)
https://www.oxforddnb.com/, accessed February 2021

Oxford English Dictionary Online (*OED*)
http://www.oed.com, accessed February 2021

Special studies

Ahmad, K. – M. Rogers
2001 "Corpus linguistics and terminology extraction". In: S.E. Wright – G. Budin (eds.) *Handbook of Terminology Management.* Vol. 2: *Application-Oriented Terminology Management*. Amsterdam: John Benjamins, 725-760.

Baker, P.
2006 *Using Corpora in Discourse Analysis.* London: Bloomsbury.

Barber, C.

1997 *Early Modern English*. Edinburgh: Edinburgh University Press.

Biber, D.

2009 "Corpus-based and corpus-driven analyses of language variation and use". In: B. Heine – H. Narrog (eds.) *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press, 159-191.

Boukhaled, M.A. – B. Fagard – T. Poibeau

2019 "The dynamics of semantic change: A corpus-based analysis". In: J. van den Herik – A.P. Rocha – L. Steels (eds.) *Agents and Artificial Intelligence. 11th International Conference, ICAART 2019, Prague, Czech Republic, February 19-21, 2019, Revised Selected Papers.* Cham: Springer, 1-15.

Brewer, C.

2010 "The use of literary quotations in the *Oxford English Dictionary*", *The Review of English Studies* 61 (248), 93-125.

2013 "*OED Online* re-launched: Distinguishing old scholarship from new", *Dictionaries: Journal of the Dictionary Society of North America* 34, 101-126.

Cabré Castellví, M. T.

1998 *Terminology: Theory, Methods and Applications*. Amsterdam: John Benjamins.

Cheung Salisbury, M.

2009 "Sarum use. The ancient customs of Salisbury", *The Journal of Ecclesiastical History* 60 (3), 582.

Coleman, J.

2013 "Using dictionary evidence to evaluate authors' lexis: John Bunyan and the *Oxford English Dictionary*", *Journal of the Dictionary Society of North America* 34, 66-100.

Considine, J.

2009 "Literary classics in *OED* quotation evidence", *The Review of English Studies* 60 (246), 620-38.

Conway, S.

2009 "Christians, Catholics, Protestants: The religious links of Britain and Ireland with continental Europe, c.1689-1800", *The English historical review* 124 (509), 833-862.

Durkin, P.

2014 *Borrowed Words: A History of Loanwords in English*. Oxford: Oxford University Press.

Gamper, J. – O. Stock

1998 "Corpus-based terminology", *Terminology* 5 (2), 147-159.

Goodland, G.

2013 "Reading Early Modern literature through *OED3*", *English Text Construction* 6 (1), 17-39.

Gotti, M.

2005 *Investigating Specialized Discourse*. Bern: Peter Lang.

Gotti, M. – S. Šarčević

2006 *Insights into Specialized Translation*. Bern: Peter Lang.

Görlach, M.

1991 *Introduction to Early Modern English*. Cambridge: Cambridge University Press.

Hadni, M. – A. Lachkar – S.A. Ouatik

2014 "Multi-word term extraction based on new hybrid approach for Arabic language", *Computer Science & Information Technology* 4, 109-120.

Hammond, M.

2020 *Python for Linguists*. Cambridge: Cambridge University Press.

Hanks, P.

2012 "The corpus revolution in lexicography", *International Journal of Lexicography* 25 (4), 398-436.

Hughes, G.

1988 *Words in Time: A Social History of English Vocabulary*. Oxford; New York: Basil Blackwell.

Isozaki, H.

2001 "Japanese named entity recognition based on a simple rule generator and decision tree learning". In: B.L. Webber (ed.) *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics, 314-321.

Jang, H. – Y. Leong – B. Yoon

2021 "TechWord: Development of a technology lexical database for structuring textual technology information based on natural language processing", *Expert Systems With Applications* 164, 114042.

Jones, H.L. (ed.)

1978 [1917] *Strabo: Geography.* Book VI. Cambridge, Mass.: Harvard University Press.

Kay, C. – K. Allan

2015 *English Historical Semantics*. Edinburgh: Edinburgh University Press.

Kilgarriff, A. et al.

2004 "Itri-04-08 the Sketch Engine", *Information Technology* 1, 105-116.

2014 "The Sketch Engine: Ten years on", *Lexicography* 1, 7-36.

Ma, L.

2011 "Clarification on linguistic applications of Fuzzy Set Theory to natural language analysis". In: *Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 26-28 July 2011, Shanghai, China*. Institute of Electrical and Electronics Engineers, IEEE Circuits and Systems Society, 811-815.

Mei, L. et al.

2016 "A novel unsupervised method for new word extraction", *Science China Information Sciences* 59, 92102.

Mills, D.

2003 *A Dictionary of British Place-Names.* Oxford: Oxford University Press.

Nevalainen, T.

2000 "Early Modern English lexis and semantics". In: R. Lass (ed.) *The Cambridge History of the English Language. V*ol. 3: *1476-1776*. Cambridge: Cambridge University Press, 332-458.

Pantel, P. – D. Lin

2001 "A statistical corpus-based term extractor". In: E. Stroulia – S. Matwin (eds.) *Advances in Artificial Intelligence. 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2001 Ottawa, Canada, June 7-9, 2001. Proceedings.* Berlin; Heidelberg: Springer, 36-46.

Pecina, P.

2010 "Lexical association measures and collocation extraction", *Language Resources and Evaluation* 44 (1/2), 137-158.

Schäfer, J.

1980 *Documentation in the O.E.D.: Shakespeare and Nashe as Test Cases*. Oxford: Clarendon Press.

Stanković, R. et al.

2016 "Rule-based automatic multi-word term extraction and lemmatization". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* European Language Resources Association (ELRA), 507-514.

Tognini-Bonelli, E.

1996 "Towards translation equivalence from a corpus linguistics perspective", *International Journal of Lexicography* 9 (3), 197-217.

Weikum, G. et al.

2012 "Big Data methods for Computational Linguistics", *IEEE Data Engineering Bulletin* 35 (3), 46-64.

Wright, S.E. – G. Budin (eds.)

2001 *Handbook of Terminology Management.* Vol. 2: *Application-Oriented Terminology Management*. Amsterdam: John Benjamins.

Xiao, R.

2008 "Theory-driven corpus research: Using corpora to inform aspect theory". In: A. Lüdeling – M. Kytö (eds.) *Corpus Linguistics: An International Handbook*. Vol. 2. Berlin: Mouton de Gruyter, 987-1008.

Address: Angela Andreani, Università degli Studi di Milano, Department of Foreign Languages and Literatures, Piazza S. Alessandro, 1, 20123 Milan, Italy.
ORCID code: orcid.org/0000-0002-5331-8915

Address: Daniel Russo, University of Insubria, Department of Human Sciences and Territorial Innovation (DiSUIT), via Dunant 7, 21100 Varese, Italy.
ORCID code: orcid.org/0000-0002-6164-1384

Token: A Journal of English Linguistics 13, 2021

Corpus stylistics, classic children's literature and the lexical field of laughter

John Corbett* & Li Li**

* *BNU-HKBU United International College*

** *Macao Polytechnic University*

ABSTRACT

This article draws upon data from the *Historical Thesaurus of English* (*HTE*) to explore the lexical field of laughter in a corpus of thirteen children's novels. The thirteen novels are all from the first 'Golden Age' of children's literature in English, namely the late 19th and early 20th century. The study takes items in the *HTE's* lexical domain of laughter and identifies the frequency, distribution and collocation of those items as they appear in the corpus. The results are discussed with reference to the ways in which different lexicalisations of laughter indicate the authorial stance towards childhood and also towards members of the communities represented in the novels. By indicating, through the representation of different types of laughter, the author's preferred moral stance towards particular individuals and groups, the novels prompt young readers to accept or challenge modes of behaviour that exemplify or threaten communal values and good citizenship. The study thus demonstrates how readers of children's literature from the 'Golden Age' are linguistically conditioned to reject negative forms of laughter and instead embrace positive forms, as they move from the undisciplined laughter of childhood to the relative restraint of adulthood.

Keywords: *Historical Thesaurus of English*, corpus stylistics, laughter, children's literature.

1. Introduction

This study combines both established and innovative lexicographical resources with text analysis software to explore the lexical domain of laughter in a small corpus of 13 novels from the first 'Golden Age' of children's literature in the 19th and early 20th centuries. The thirteen novels

DOI: 10.25951/4821

are selected from a rich period in which, critics generally agree, children's authors balanced their moral and educational aspirations with a desire to entertain children in their own terms (see Carpenter 1985; Darton 1982 [1932]; Sorby 2011). We were therefore interested in exploring, stylistically, in the novels the frequency and distribution of a set of lexical items that indexed both moral stance and pleasure. Given that laughter can represent a range of expressive and relational stances, from the spontaneous outburst of pleasure to a sneering indication of superiority, we have investigated the frequency and distribution of a set of items in the lexical field of laughter, as they appear in a number of 'Golden Age' novels. We address the following questions:

- Which items in the lexical field of laughter appear in the novels?
- What is the distribution of these items across the novels?
- What do the collocations of these items tell us about the conceptual domain of laughter, as it is represented in the novels?
- How do the choices made from the lexical domain of laughter manage the readers' stance towards the characters in the novels, and the characters' relations with each other?

To answer these questions, we began by consulting the *Historical Thesaurus of English* (*HTE*; see Kay et al. 2009) in order to identify the expressions related to laughter that were available during the first 'Golden Age' of children's literature. We then used the text analysis software, AntConc (Anthony 2019), to determine which lexical items from that set of expressions were actually used in the novels, to calculate their frequency, distribution, and statistical significance, and to explore how they were modified. The corpus-informed stylistic approach that we take to the study of laughter in a corpus of children's novels is quantitative, and it can be understood as a type of 'distant reading' (Moretti 2005).

2. The lexical field of laughter

The availability over the past decade of the *Historical Thesaurus of English* has made possible certain types of investigation into semantics and stylistics that were hitherto impossible (see Anderson et al. 2016; Busse 2012). We consulted the *HTE* to identify a set of words within the lexical field of laughter that were current during the first 'Golden Age' of writing for children. Then we used text analysis software to find the frequency and distribution of the various expressions within the corpus. Finally, collocations of *laugh/laughter* were analysed to discover the ways in which acts of laughing and laughter are conceptualised in the novels. The collocations show that certain concepts related to laughter, such as spontaneity, control, and sincerity, are salient in the corpus. The analysis suggests that, in these novels, unrepressed, communal laughter is an index of unconstrained youth, and that learning to identify the 'proper' kinds of laughter is a rite of passage from childhood to adulthood.

The methods parallel other corpus stylistics investigations, such as Oster's (2010) use of corpora to explore the linguistic expression of emotions. Based on the historical dictionaries of English, principally the *Oxford English Dictionary* (*OED*), the *HTE* classifies the lexical resources of English into 250,000 discrete concepts, divided first into The World, The Mind and Society. These concepts include the verb and noun *laugh*, the noun *laughter,* and all the expressions in English that are semantically related to them, such as *giggle, titter, sneer, chuckle, chortle*, and so on. The *HTE* also indicates the chronology of the development of the lexical domain. Both *laugh* and *laughter* are recorded as far back as the early Old English period (from OE *ahliehhan, hleahter,* etc.). The *OED* suggests that the etymology is imitative, and that the word is Indo-European in origin. Although the form of the present-day word *laughter* dates from Old English, its meaning has changed. While in OE texts, *laughter* could signify a single instance of laughter, and appear as a plural (*hleahtres*), from at least the 16th century, the verbal form *laugh* began also to be used as a countable noun indicating an instance of laughter, and the form *laughter* was increasingly reserved for use as a mass noun with a more generalized meaning.

Specific ways of laughing in English are more recent. The *OED* suggests that *giggle* is echoic, but its use as a verb is not recorded until 1509 and a nominal usage is not recorded until 1611. The expression *chortle* – apparently a blend of *chuckle* and *snort* – is not recorded until 1871, when Lewis Carroll coined it in the poem 'Jabberwocky.' Again, the verbal use precedes the nominal use, which is not recorded in the *OED* until 1903. The expression, now reasonably common, appears only once in our corpus, in the second of the 'Alice' books, *Through the Looking Glass*, in which the poem, 'Jabberwocky,' first appeared.

While the lexical domain of laughter contains other forms of words – the adverb *laughingly* is recorded from 1475 and *sniggeringly* from 1886, we focus here, for reasons of space, largely on verbal and nominal forms and senses that indicate types of laugh/laughter that are current during the period of our corpus (1865-1911). From the evidence of the *HTE*, in this lexical domain, verbal forms and senses usually predate the nominal ones. Certain expressions arose in English and died out before the period of the corpus (e.g. *unlaugh,* 'to reverse the laughing process,' is sporadically recorded only in 1532 and 1637), while others appear later (e.g. *hoot* in the sense of *laugh* is not recorded until 1926, and *laugh-in*, an event characterised by laughter, is not recorded until 1968). As we have already noted, although some words endure, their senses change; thus from 1598-1823, the word *chuckle* is recorded with the sense of laughing convulsively or immoderately, but thereafter its meaning weakens or narrows to the sense of laughing quietly and with contentment, a sense first recorded in 1803.

The selection of search items for this study was motivated by a concern to understand how laughing and laughter are deployed stylistically in the corpus. We are thus less concerned with whether *laugh* is used in the texts as a verb or a noun, and more with whether, when and why characters (give a) *laugh, giggle, chuckle, snort,* or, indeed, *chortle*. We are also concerned with how these expressions are modified, that is, whether, when, and why characters might give, for example, a brave, honest laugh, or a cowardly, hollow laugh. We are concerned, in short, with how laughter relates to the moral universe of the texts.

The intransitive verb, *laugh,* and the related noun, *laughter,* are both coded 02.04.10.11 in the *HTE*. This code indicates that the lexical domain of *laugh/laughter* has been categorized under 02 The Mind > 04 Emotions > 10 Pleasure > 11 Laughter. This coding is evidently based on the sense of *laugh/laughter* as a vocal outburst that is expressive of pleasure; there are other senses and homonyms that do not concern us here, such as the northern English and Scots use of *a laughter* referring to the total number of eggs laid by a hen or another bird, e.g. 'A hen lays her laughter, that is, all the eggs she will lay that time.'[1] These unrelated senses are excluded from our analysis.

Further and finer *HTE* codings indicate relevant subcategories of *laugh/laughter*. The search items selected for inclusion in our analysis are based on the lists in Appendices 1 and 2. We have excluded from the lists in the Appendices those items that were current in the period, according to the *OED/HTE*, but which do not appear in our corpus, such as *cachinnate* 1824- 'to laugh loudly/coarsely'. The potential search items in Appendix 1 are based solely on the intransitive uses of the verb – the transitive list in

[1] "laughter, n.2" (*OED Online*).

the *HTE* brings up only one further possible item, namely *guff* (1865-) in the sense of 'utter/express with loud/coarse laughter' and this item, like the similar *guffaw*, is absent from our corpus. The nominal forms in Appendix 2 largely repeat exactly the word forms of the verbs, with a few exceptions, e.g. the verb *convulse* corresponds to *convulsion*(*s*) in the sense of 'a fit or fits of laughter.'

There are a few points worth noting at the outset about this lexical domain. Since the action of laughing involves, to quote part of the first *OED* definition of *laugh*, 'the spontaneous sounds and movements of the face and body usual in expressing joy, mirth, amusement, or (sometimes) derision,'[2] it is not surprising that many of the items in the lexical domain of laughter (including *laugh* itself) are imitative or echoic, for instance, *roar, snort, haw-haw, tee-hee, giggle, titter, cackle.* Although laughter is a spontaneous expression of pleasure, arguably only *chuckle* is now used to express moderate contentment. The animal nature of laughter is salient in expressions such as *roar, snort, whinny, cackle,* and *horse-laugh.* The idea of a spontaneous outburst that cannot be controlled is evident in expressions like *die with laughter, split the sides, laugh oneself sick/silly*, *break up, convulsion*(*s*), etc. The lack of control associated with certain forms of laughter may be associated with foolish laughter, for example, *giggle,* or *titter*. Furthermore, although laughter is categorised in the *HTE* primarily in relation to pleasure and mirth, an element of derision is highlighted in numerous expressions, such as *snicker, snigger, cackle, laugh in one's sleeve.* Finally, there is evidently some overlap and fuzziness in the use of these expressions. For example, the word *sneer* is categorised within 'foolish laughter,' but it can also mean 'derisory laughter'; and different senses of *chuckle* mean that it can be considered either as a type of 'snigger' or a subcategory by itself.

Despite these instances of fuzziness and overlap, the lexical domain of *laugh/laughter*, as shown in the *HTE*, gives a reasonably clear indication of the extent and boundaries of the concept in English: laughter is a physical, usually vocal, outburst of mirth that might indicate pleasure or derision; it can be uncontrolled and immoderate, and may be suggestive of an animal nature or foolishness. These are the conceptual traits of laughter as described in reference works like dictionaries and thesauri. However, *laugh/laughter* can also be modified in context to suggest a broader range of characteristics, and to analyse this phenomenon further, we need to turn from lexicographical works of reference to a corpus.

[2] "laugh, v." (*OED Online*).

3. The corpus of 'Golden Age' novels

The present study explores the uses of laughter in a small corpus of thirteen children's novels in what is generally considered the first 'Golden Age' of children's literature in English. Harvey Darton and Humphrey Carpenter suggest that this 'Golden Age' begins with Lewis Carroll's *Alice's Adventures in Wonderland* (1865) and draws to a close with A.A. Milne's stories of Winnie-the-Pooh, the last of which was *The House at Pooh Corner* (1928). The 'Golden Age' canon consists of texts written specifically for children that are characterised by an emphasis on entertainment as much as or rather than moral instruction (Sorby 2011), and a new willingness to view the world from the perspective of a child, like Alice, or Peter Pan, or a child-surrogate, such as Beauty in *Black Beauty*, or Buck in *The Call of the Wild*.

Certainly, the texts of the period acknowledge the complexities a child faces when negotiating and eventually entering the world of adults. The specific question that the present study raises is the role of laughter in making that transition. The thirteen novels in the corpus were downloaded in digital form from the Project Gutenberg website[3] and edited to remove extraneous matter. The novels are listed in Table 1. They were not selected as necessarily being representative of all writing for children in this period; we wished simply to be able to compare the uses of laughter in any single novel of the period with the uses of laughter in a reasonable number of other novels of the same period. Lewis Carroll's *Through the Looking Glass* was added to the corpus when we realised that there are no occurrences whatsoever of expressions relating to *laugh/laughter* in *Alice's Adventures in Wonderland,* a fact that is interesting, in itself, and which is discussed below. The titles, authors, dates of publication and provenance (UK/USA) are given, with the word count for each novel. The provenance is given as there might be a preference for a particular form (e.g. *snicker/snigger*) in American or British English.

The lexical field of laughter affords the authors of these novels a range of possible expressions with more or less specific senses by which they can portray characters and their actions, and thus indicate the stance of the characters towards themselves, others, and events in the world. The expressions within the domain of laughter can themselves be modified, particularly the less specific, superordinate term, *laugh,* which is general in meaning, no matter whether it is used as a noun or a verb. The action of

[3] *Books in Children's Literature* (Project Gutenberg).

laughing can be expressed and modified either by using a verb, i.e. *she laughed*, or by using a delexicalized or 'light verb' (that is, verb whose meaning is reduced and the nature of the action is conveyed by the grammatical object, as in *she gave a laugh/she had a laugh*, etc).

Table 1. The Corpus

Text No.	Code	Title	Author	Date	Number of words (tokens)	Provenance
1	AW	*Alice's Adventures in Wonderland*	Lewis Carroll	1865	10,021	UK
2	LG	*Through the Looking Glass*		1871	30,618	
3	LW	*Little Women*	Louisa May Alcott	1868	191,196	USA
4	WKD	*What Katy Did*	Susan Coolidge	1872	51,129	USA
5	BB	*Black Beauty*	Anna Sewell	1877	60,848	UK
6	TI	*Treasure Island*	Robert Louis Stevenson	1883	70,425	UK
7	WWO	*The Wonderful Wizard of Oz*	L. Frank Baum	1900	39,888	USA
8	CW	*The Call of the Wild*	Jack London	1903	32,365	USA
9	RSF	*Rebecca of Sunnybrook Farm*	Kate Douglas Wiggin	1903	76,090	USA
10	WW	*The Wind in the Willows*	Kenneth Grahame	1908	60,754	UK
11	PPW	*Peter Pan and Wendy*	James M. Barrie	1911	48,178	UK
12	SG	*The Secret Garden*	Frances Hodgson Burnett	1911	83,164	UK
13	DD	*The Story of Doctor Dolittle*	Hugh Lofting	1920	27,570	UK
Total number of tokens in the corpus					782,246	

The use of 'light' verbs allows for a wide range of modifications, e.g. *she gave a hearty/delicate/sinister/hollow laugh,* etc. The generality of *laugh* is evident when one attempts to substitute it with one of its subcategories; one can hardly give a **hearty giggle, *delicate roar, *sinister guffaw* or **hollow titter.*[4] The more restricted senses of the expressions in the subcategories constrain the ways in which they are modified, e.g. *hearty roar, foolish giggle.* While the description of a character as *chuckling, sneering* or *giggling* may be taken to be indicative of a character's personality, then, the act of laughing is, by itself, less revealing. It is therefore necessary to look at the immediate contexts in which the terms *laugh, laughter* and their related forms, occur, and consider how they are modified.

From the lexical domain of laughter, as categorised by the *HTE*, we have selected as search items a set of lemmas that occur in our corpus (Table 2). The lemma *laugh** searches for all forms of the verb, *laughs/laughed/*etc. plus the nouns *laugh* and *laughter*. A number of relatively common expressions, such as *guffaw, shake with laughter, tee-hee* do not occur in any of the novels in our corpus, and so they have been omitted from the table. The expression *split my sides* is uttered as an oath by a pirate in *Treasure Island*, but it does not seem to be related to laughing. Table 2 summarises the selected search items plus their senses. Further details about these expressions are given in Appendices 1 and 2.

Our exploration of the lexical domain of laughter in the corpus of thirteen novels consisted of a series of searches using the text analysis tool, AntConc version 3.5.8 (Anthony 2019). The first set of searches addressed the frequency and distribution of the *HTE* search items in each of the novels, and measured the 'keyness' of the items, that is, whether their frequency of occurrence in any given novel is statistically higher or lower than might be expected from an analysis of the corpus as a whole. The findings of these searches indicate the degree to which each novel draws upon the lexical domain of laughter – and, if so, where. *Alice's Adventures in Wonderland* is remarkably devoid of laughter, as we have already noted, while *Little Women* and *The Secret Garden* overflow with it. The occurrences in the other novels fall in between.

[4] As is conventional, we indicate an unacceptable usage by placing an asterisk before it, e.g. **hearty snigger*. An asterisk at the end of an item, such as *laugh**, indicates that it is a lemma, or 'wild card' used in corpus searches to identify instances not only of *laugh* but also of *laughs, laughed, laughing, laughter*, etc.

Table 2. Search items from the lexical domain of *laughter*

Search item	Meaning
laugh*	the expression of pleasure or derision through sounds, bodily movement and/or facial expressions; an instance of such expression; to express pleasure or derision in this way.
roar* snort* ha ha*	(give) an outburst of loud laughter
die* with/of laughing convulsion*	(give) an outburst of immoderate or wild laughter
giggle* titter*	(give) a giggle
sneer*	(give) a foolish or mocking laugh
whinny*	(give) a laugh in the manner of a horse
snicker* snigger* cackle*	(give) a snigger
chuckle* chortle*	(give) a chuckle

The second set of searches focuses on the lemma *laugh** and considers its lexical contexts in the corpus. The lexical company kept by the forms of *laugh** in each of the texts indicates whether laughter and laughing are semantically positive or negative concepts in the novels. Finally, the discussion reviews the findings and considers how each of the novels draws upon the conceptual field of laughter as part of the moral universe of these 'Golden Age' narratives.

4. Frequencies, distributions and keyness

Appendix 3 shows the raw frequency of each search item in the corpus, that is the number of times the lemma appears in each text, plus the normalised frequency, which refers to the number of occurrences per 1,000,000 words. Normalising the frequencies allows for comparison between texts that differ in length. The shortest text contains 10,021 word tokens (*Alice's Adventures in Wonderland*) and the longest contains 191,196 word tokens (*Little Women*). Normalised frequencies are used as the basis for the charts shown in the present section of this article; figures have been rounded to the nearest whole number, for clarity of presentation.

Some broad preliminary observations can be made on the basis of this data. Table 3 shows a chart of the normalised frequency of *laugh** in the individual texts in the corpus. As we have already observed, it is a curious fact that no items from the lexical domain of laughter appear in *Alice's Adventures in Wonderland*. Only when she goes through the looking glass in the second novel does Alice, along with several other characters, laugh. By contrast, in *The Secret Garden* and *Little Women*, one or another of the forms of *laugh/laughter* appears over 1200 times per million tokens. As the chart shows, the other novels contain a varying frequency of occurrences of the forms.

Table 3. Normalised frequency of *laugh** in the novels (per million words)

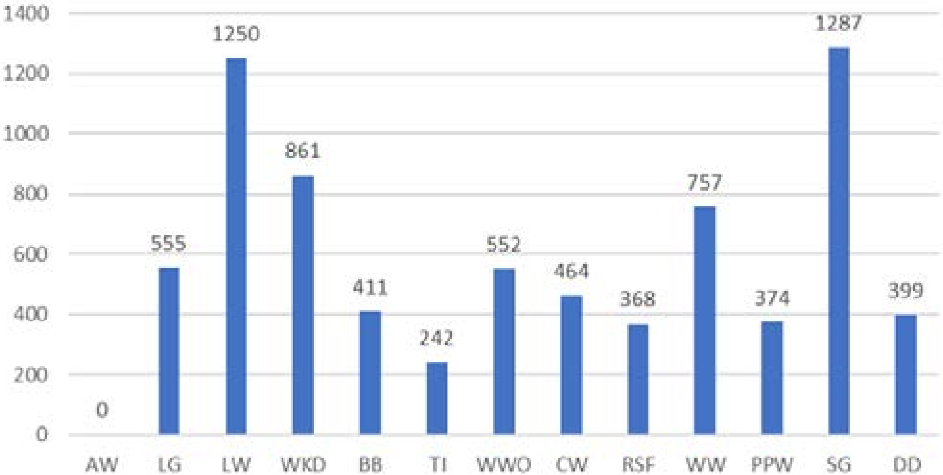
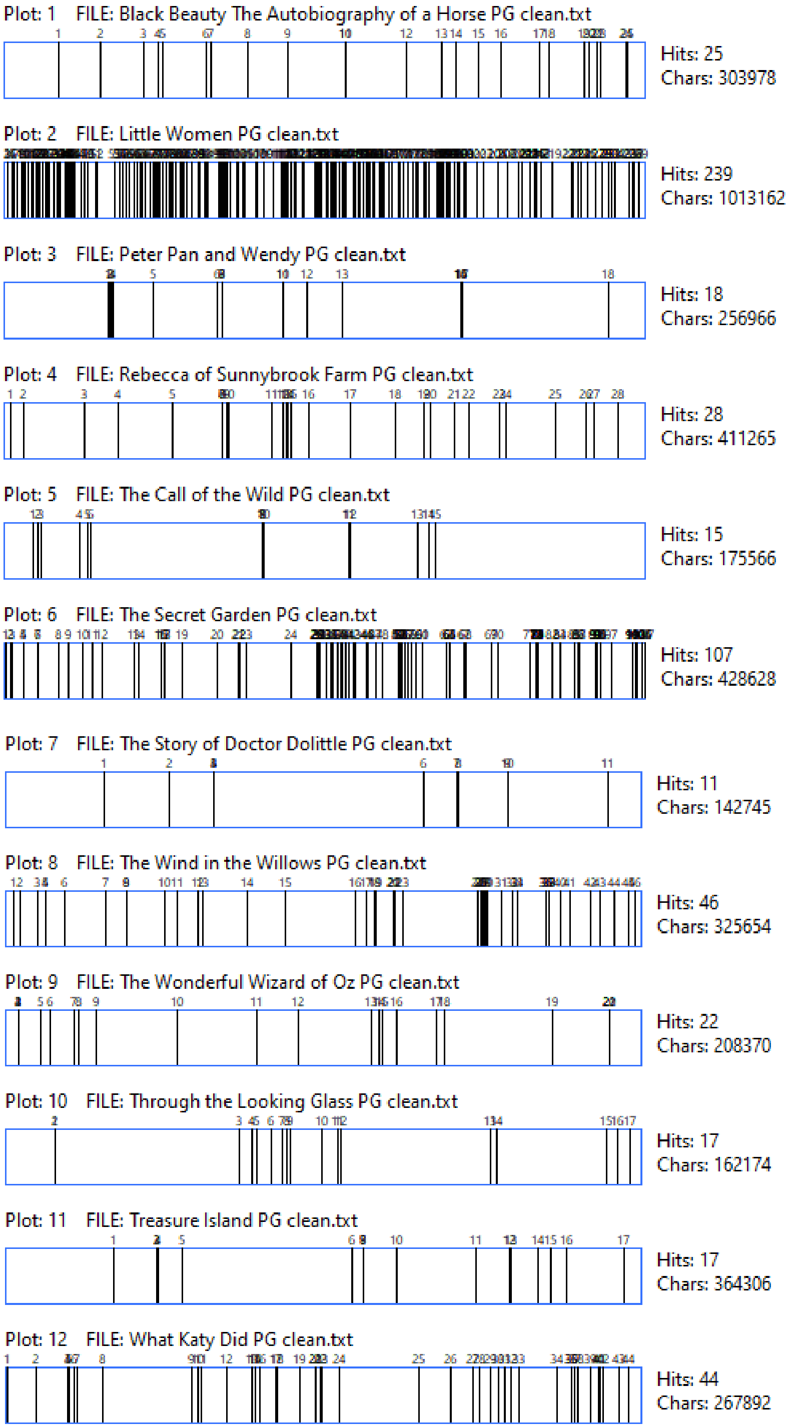


Table 4 shows what AntConc labels the 'concordance plot' of the lemma *laugh** in the twelve novels in which forms of *laugh/laughter* appear (*Alice in Wonderland* is not included, as there are no occurrences). This plot shows the distribution of the expressions of laughter in the novels. This table adds some further detail to the frequency data; for example, while *The Secret Garden* has more occurrences than *Little Women*, the distributions differ to some extent.

Table 4. Concordance plots of laugh* in the corpus



Plot: 1 FILE: Black Beauty The Autobiography of a Horse PG clean.txt
Hits: 25
Chars: 303978

Plot: 2 FILE: Little Women PG clean.txt
Hits: 239
Chars: 1013162

Plot: 3 FILE: Peter Pan and Wendy PG clean.txt
Hits: 18
Chars: 256966

Plot: 4 FILE: Rebecca of Sunnybrook Farm PG clean.txt
Hits: 28
Chars: 411265

Plot: 5 FILE: The Call of the Wild PG clean.txt
Hits: 15
Chars: 175566

Plot: 6 FILE: The Secret Garden PG clean.txt
Hits: 107
Chars: 428628

Plot: 7 FILE: The Story of Doctor Dolittle PG clean.txt
Hits: 11
Chars: 142745

Plot: 8 FILE: The Wind in the Willows PG clean.txt
Hits: 46
Chars: 325654

Plot: 9 FILE: The Wonderful Wizard of Oz PG clean.txt
Hits: 22
Chars: 208370

Plot: 10 FILE: Through the Looking Glass PG clean.txt
Hits: 17
Chars: 162174

Plot: 11 FILE: Treasure Island PG clean.txt
Hits: 17
Chars: 364306

Plot: 12 FILE: What Katy Did PG clean.txt
Hits: 44
Chars: 267892

The apparent degree of difference in the distributions is perhaps exaggerated by the fact that the concordance plots are normalised in size, and *Little Women* contains more than double the number of word tokens that *The Secret Garden* does. Even so, in the latter novel, the occurrences of laughter come in bunches or waves, while in the former, laughter saturates much of the novel, tailing off towards the conclusion.

A final measure of the frequency of *laugh/laughter* in the corpus is 'keyness,' which is a measure of the extent to which the frequency of occurrence of the forms in one text or set of texts is statistically more or less what might be expected when considered with reference to a larger corpus. The measure of 'keyness' of *laugh/laughter* thus indicates the extent to which it is being used *unusually* frequently or infrequently in a particular text, compared to that in other texts. Obviously, much depends on the reference corpus to which the individual novel, in our case, is being compared. For the purposes of this study, we have taken the thirteen novels as a whole to be the reference corpus against which the individual novels are compared. Since AntConc does not lemmatise words in a key word search, we have focused on measurements of the keyness of three main word-forms, namely *laugh, laughed,* and *laughing*. There are different statistical ways of calculating keyness; in this study we have used a log likelihood measure. With this measurement, a value above +6.63 is generally considered to be statistically significant; that is, if the keyness exceeds that value, we can confidently assert that the word-form is used unusually frequently in the given text. A negative value suggests that, where it is used, the form is unusually infrequent. Again, *Alice's Adventures in Wonderland* has not been given a score as no forms of the word appear in the text.

Table 5. Keyness values for *laugh** in each of the novels

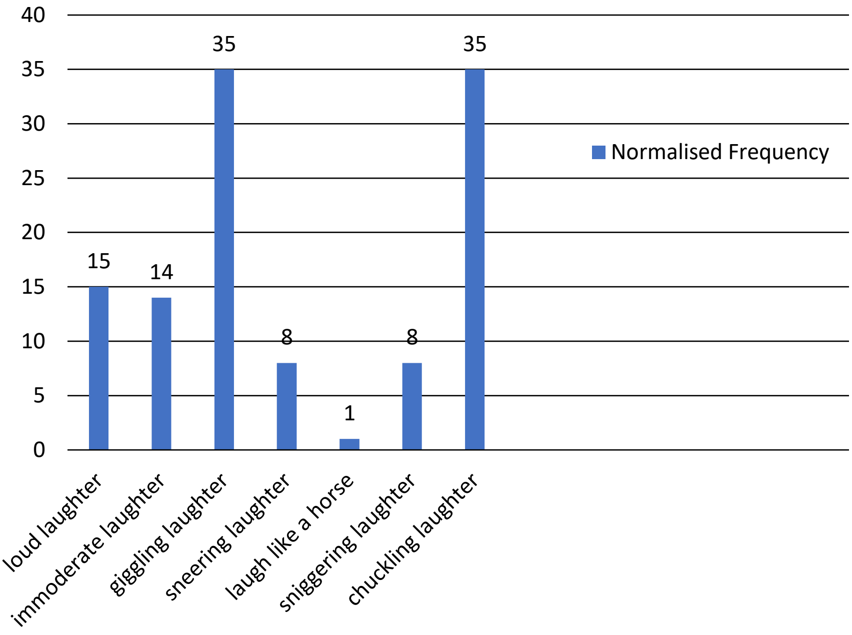
	AW	LG	LW	WKD	BB	TI	WWO	CW	RSF	WW	PPW	SG	DD
Laugh	–	–0.11	+33.24	–0.02	–11.54	–5.75	–1.04	–5.18	–16.58	–5.06	–1.44	+3.65	–1.96
Laughed	–	–1.16	+11.88	+0.24	–3.12	–12.53	-0.51	+0.31	–0.63	–1.42	–7.39	+7.35	–2.93
Laughing	–	–0.06	+1.71	+1.18	+0.06	–13.52	-0.78	–1.7	–11.84	+5.24	–7.38	+8.7	0

The main finding in the key word calculation is that the main lexical forms associated with *laugh* are very unusually frequent in *Little Women*, their keyness scores being considerably higher even than those of *The Secret Garden.* By comparison, there is a relatively infrequent use of the forms of *laugh* in *Rebecca of Sunnybrook Farm,* and *Treasure Island*, where keyness values are negative. In most of the other novels, the values suggest neither a significantly frequent nor significantly infrequent use of the forms with respect to the corpus as a whole.

The frequency of occurrence of the forms of the lemma *laugh** by far exceeds that of any of the other items in the lexical domain of laughter. As can be seen in Appendix 3, there are only sporadic uses of the other items in the lexical domain. *Little Women* again has the widest range of expressions within the lexical domain (*ha ha; die of laughing, convulse/convulsion, giggle* and *chuckle*) but even the most common of these (*convulse/convulsion*) has a normalised frequency of only 36.6 per million words, compared to 1250 per million for *laugh**.

A close study of Appendix 3 prompts a number of noteworthy observations: within our corpus, *Treasure Island, Peter Pan and Wendy* and *The Call of the Wild* have a near monopoly on *sneer**; while *The Wind in the Willows* and *The Secret Garden* have a high normalised frequency of *chuckle**. However, the frequencies are too small for an analyst to make much of. It is more illuminating to group the lexical items in the corpus as a whole according to the semantic categories suggested by the *HTE* (Table 6).

Table 6. Types of *laughter* in the corpus – normalised frequency (per million words)



Admittedly, when categorising some of the items used in the calculations in Table 6, there is a degree of subjectivity on the part of the analyst which affects the results when the occurrences are so low. Some of the items in the lexical domain of laughter are associated with animal or bird sounds (*roar, snort, whinny, cackle*) and it is a staple of children's fiction that animals can be major characters. Thus, in *The Wizard of Oz*, the Cowardly Lion roars (but never with laughter), Beauty in *Black Beauty* snorts and whinnies (but only once with joy), and Polynesia, the parrot in *The Story of Doctor Dolittle*, cackles. Concordance analysis has been used to disambiguate some of these usages, and, unless there is an explicit indication that the roaring, snorting, or whinnying was an expression of pleasure or derision, then the item was excluded from the figures shown.

Overall, what the present section indicates is how the general conceptual domain of laughter is lexicalised in English and how the members of the lexical domain are realised in a corpus of children's novels. So far, we have seen that the lexical domain of laughter can be completely ignored (as in *Alice's Adventures in Wonderland*) or it can be so frequently utilised (as in *Little Women*) as to be an obvious thematic element. The children's novels also show a relatively high frequency of loud and immoderate laughter, giggling and chuckling – there is clearly a concern with laughter as something irrepressible and uncontrollable, or as a signal of quiet contentment. However, expressions indicating other forms of laughter – sneering, sniggering or horse-like – are relatively less in evidence.

The analysis so far has indicated the relative presence and absence of members of the lexical domain of laughter in the English language in thirteen novels from the first 'Golden Age' of children's literature; we now turn to a more detailed examination of the uses of laughter in the texts.

5. Collocates and concordances

The collocation of a word is, to quote John Firth, "the company it keeps" (Firth 1957: 11; see also Bartsch 2004; Sinclair 1991). Those who study semantic preference suggest that the habitual presence of a frequently occurring collocate can impact on the meaning of a word; for example, they argue that since *regime* frequently collocates with modifiers such as *brutal* and *repressive*, the concept of a *regime*, in contrast to, say, that of a *government*, begins to acquire negative connotations. The implications drawn from this claim can be contentious (cf. Hunston 2007; Stewart 2010), but the fact remains that habitual co-occurrence among words may make salient certain aspects of their potential meaning. In a study of the meanings and uses of the expression *Irish* in a corpus of British parliamentary discourse, Corbett (2021) argues that diachronic corpus stylistics can illustrate the dynamics by which the association of a particular term with a positive or negative attitude can be reinforced or challenged.

We have seen in the foregoing sections that the members of the lexical domain of *laugh/laughter* enable English speakers to make salient the particular, embodied, and relational aspects of laughing and laughter: its volume, its spontaneous and uncontrollable nature, its animal-like qualities, its potentially derisive import, and whether or not it is concealed. We would expect habitual collocates of *laugh/laughter* to express similar meanings, and to extend them. That is, collocations such as *burst out laughing* express the spontaneous, uncontrolled nature of the behaviour, while *laugh at/with* express the relational meanings of mockery or solidarity and empathy. Further collocations such as *frank/hollow laughter* extend the meanings of laughter to encompass concepts like sincere/insincere laughter, for which no single word is available in English.

There are two main ways of analysing co-occurrence using corpora. The researcher can look at a concordance line, that is, a stretch of text that spans a given number of words on either side of the search item or node. This technique is used when the analyst wishes to identify and interpret patterns manually, and we do this with some selected data below. However, when dealing with a large amount of textual data, programs like AntConc also search automatically for frequently occurring items on either side of a given node. The results of such searches depend on the span selected, and they usually show two types of result: the *frequency* of co-occurrence, which is self-explanatory, and the *strength* of the collocation. The latter value is a statistical measure of the likelihood or probability of co-occurrence, which can be calculated in a number of ways; however, no matter how it is calculated, the probability value depends on the overall collocational behaviour of items in the corpus. Thus, for example, the words *hollow* and *honest* might both modify *laughter* only once in the corpus. Their frequency of co-occurrence would obviously be equal. However, if *honest* modifies a number of other words in the corpus, while *hollow* is largely reserved for *laughter*, the strength of collocation between *laughter* and *hollow* will be greater than that between *laughter* and *honest*. The same statistical calculation often means that a less frequently occurring collocate has a higher collocational strength. The analyst thus needs to pay attention to the values of both frequency and strength of collocation, since an understanding of both is necessary in order to understand how collocates are behaving in a corpus. The measure of collocational strength used in the analyses below is Mutual Information (MI) and it is conventionally assumed that an MI score of 3 or above indicates that there is a statistically strong bond between the items in question.

Tables 7 to 10 show the results of several collocation searches for members of the lemma *laugh** in the corpus as a whole. Table 7 is given largely to illustrate the principle of collocation by showing ways of describing instances of laughter in the corpus, through a search for expressions that occur one item to the left of the phrase *of laughter.* The collocates listed indicate, again, the nature of laughter in the corpus, namely that it is sudden, involuntary and noisy. It can be violent (*explosions, convulsions*) but can have pleasant associations with, say, the pealing of bells. Of the 10 collocates listed, the first 5 have a statistically strong association (MI of 3 or above) with *of laughter*. However, of these 5, only *peals* occurs more than once in the corpus – and both times in *Little Women*, which is also the source of the singular occurrence, *peal of laughter.*

Tables 8 to 10 show a number of further searches that explore other collocational aspects of the lemma *laugh**. In these searches, a stop-word list has been used to exclude common grammatical items from the results, and the spans, though narrow, are intended to identify salient modifiers, mainly adjectives and adverbs. Table 8 shows collocates occurring one word before *laughter*, Table 9 shows adjective collocates occurring one word before *laugh*, and Table 10 shows adverbial collocates occurring one word after *laughed*. All are ranked according to MI, and only those MI values above 3 are shown in the tables.

As expected, the collocates of *laugh/laughter* confirm what we know about the basic nature of the conceptual field from a study of members of the lexical domain as represented in the corpus of children's writing (e.g. the volume of laughter can be low or uproarious, and the quality of laughter can be pleasant or disagreeable), but they also add further detail.

The collocates describing *laugh/laughter* in these tables fall into certain positive or negative thematic groups, as shown in Table 11, which groups the collocates that are listed in Tables 8-10 according to their semantic themes.

Table 7. Collocates immediately to the left of of *laughter*, sorted by MI

Rank	Frequency	MI	Collocate
1	2	5.7	peals
2	1	4.7	explosions
3	1	4.7	convulsions
4	1	4.1	shouts
5	1	4.1	explosion
6	1	2.9	peal
7	1	2.5	shrieks
8	3	1.2	burst
9	1	0.8	scream
10	1	-2.9	full

Table 8. Collocates immediately to the left of *laughter*, sorted by MI

Rank	Frequency	MI	Collocate
1	1	11.0	uncontrollable
2	1	11.0	convulsive
3	1	10.7	boisterous
4	1	10.3	mocking
5	1	9.9	suppressed
6	1	8.7	childish
7	1	8.4	boyish
8	1	7.7	hollow
9	1	7.5	careless
10	1	7.4	honest
11	1	5.4	secret
12	1	5.0	low

Table 9. Adjective collocates immediately to the left of *laugh*, sorted by MI

Rank	Frequency	MI	Collocate
1	1	10.9	sardonic
2	1	10.9	irresistible
3	1	10.9	heartier
4	1	9.3	sonorous
5	1	8.6	reckless
6	1	8.1	mocking
7	4	7.9	haughty
8	1	7.7	mellow
9	1	7.7	suggestive
10	1	7.7	hearty
11	1	6.7	hoarse
12	1	6.4	disagreeable
13	3	6.2	merry
14	1	6.0	frank
15	1	5.9	winged
16	2	5.3	comfortable
17	3	4.9	short
18	1	4.8	jolly
19	1	3.4	pleasant

Table 10. Adverb collocates immediately to the right of *laughed*, sorted by MI

Rank	Frequency	MI	Collocate
1	1	9.7	uproariously
2	1	9.7	harshly
3	4	8.4	outright
4	1	9.1	hysterically
5	1	7.5	noiselessly
6	1	7.2	scornfully
7	4	7.0	aloud
8	3	6.7	heartily

Table 11. Modifying collocates of *laugh** sorted thematically

Theme	Collocate	
	positive	negative
Control		hysterically reckless convulsive uncontrollable boisterous irresistible careless
Honesty/sincerity/ transparency	outright honest aloud frank	suppressed hollow secret suggestive
Engagement	hearty/heartier/heartily	haughty
Derision	sardonic mocking scornfully	
Maturity		childish boyish
Volume	uproariously sonorous	low hoarse noiselessly
Contentment	comfortable mellow	
Pleasure	pleasant jolly merry	disagreeable harshly
Duration		short

5. From distant to close reading

Up to this point, then, we have selected a group of expressions related to laughter, drawing upon the lexicographical resources of the *HTE*, and we have used this group as the basis for a number of corpus searches. The findings reported in the sections above show the frequency and distribution of members of the lexical set throughout our corpus, and further searches of the collocates of *laugh*(*ed*) and *laughter* have shown how the items are modified in the novels. In this section, we offer a number of interpretations that these findings would support, with extracts from the novels. Effectively, this section, then, marks a shift from distant to closer reading.

Insofar as patterns of lexical frequency and distribution and the choice of collocates suggest themes, the findings set out in the previous sections are highly suggestive. The collocates of *laugh** in Table 11 accord with the earlier tables in showing laughter in the novels as indicative of a lack of control and care. This attitude towards laughter is not necessarily negative: if children's literature of the 'Golden Age' is characterised in part by adult nostalgia for the carefree, unrepressed period of youth (when laughter might be 'childish' or 'boyish'), then convulsive, even hysterical laughter can be indicative of a stage in life when pleasure may be unconstrained and undisciplined. This view of laughter is made explicit in *The Secret Garden*:

(1) It seemed actually like the **laughter** of young things, the uncontrollable **laughter** of children who were trying not to be heard but who in a moment or so–as their excitement mounted–would burst forth.

A nostalgic yearning for a time of unconstrained pleasure is most evident in *Little Women*, where laughter is frequently a marker of communal pleasure and solidarity, and it can often send the person who laughs into fits. Laughter is, for example, one of the phenomena that mark holidays such as the annual apple-picking and Christmas:

(2) Everybody **laughed** and sang, climbed up and tumbled down. Everybody declared that there never had been such a perfect day or such a jolly set to enjoy it, and everyone gave themselves up to the simple pleasures of the hour as freely as if there were no such things as care or sorrow in the world.

(3) As Christmas approached, the usual mysteries began to haunt the house, and Jo frequently **convulsed** the family by proposing utterly impossible or magnificently absurd ceremonies, in honor of this unusually merry Christmas.

The thematic elements of honesty and engagement are more evenly balanced between positive and negative connotations, and they suggest disciplined laughter as a moral good: *suppressed* laughter is counterbalanced by *outright* laughter, *hollow* laughter by *honest* laughter, *haughty* laughter by *hearty* laughter and so on. The social values attached to the positive connotations (e.g. *honesty, heartiness*) can be seen as educative for youthful readers as they mature from childhood into adulthood: since laughter is spontaneous and not subject to control, it functions to reveal aspects of moral character, good and bad.

The theme of derision involves the collocates *sardonic, mocking, scornfully* as well as lexical items such as *sneer.* These terms represent a relational aspect to laughter, the fact that people can laugh derisively or contemptuously at others and at events. These acts of laughter signify the superior attitude that derives from the person who is laughing having power or moral authority. The expression of such power and authority through laughter may index flawed character or villainy (unsurprisingly, the pirates in *Peter Pan and Wendy* and *Treasure Island* sneer), although it is equally possible for a character, such as the doctor in *Treasure Island*, to sneer at a pirate, or a parent to laugh scornfully at a child's perceived foolishness. Peter Pan can also sneer at the laws of nature and, in the same novel, the mocking laughter of mermaids can be directed at the inadequacies of those confined to land.

Other positive and negative collocates of laughter indicate character through engagement (*hearty* versus *haughty* laughter), volume of laughter (*uproarious* versus *low*), degree of contentment (*mellow* laughter) and whether or not it signifies or causes pleasure (*disagreeable/harsh* versus *pleasant/merry* laughter). Where binary choices are available, youthful readers are socialised through their reading to recognise positive and negative character traits by the description of styles of laughter produced in the texts. While, in *Peter Pan and Wendy*, laughter is not particularly frequent, its moral nature is emphasised. One of the tasks Wendy gives to her younger brothers to remind them of home is to describe their parents' laughter. Moreover, when Hook devises a plan to lure the Lost Boys to their doom by using a cake as bait, his true nature is revealed through his laughter:

(4) 'They will find the cake and they will gobble it up, because, having no mother, they don't know how dangerous 'tis to eat rich damp cake.' He burst into **laughter**, not hollow **laughter** now, but honest **laughter**. 'Aha, they will die.'

The sinister import of this exchange is intensified by the fact that Hook's shift from insincere to sincere laughter coincides with his acknowledgement that the plan will lead to the children's death.

The major counter to our argument that the presence of laughter in children's literature socialises youthful readers into a recognition and adoption of adult social values is, of course, *Alice's Adventures in Wonderland*, in which there is no lexical trace whatsoever of any kind of laughter at all. Given that the novel is comic, this fact is perhaps surprising. However, it could be argued that laughter in this novel is displaced from the text to the reader, who is invited to observe Alice and her curious encounters in Wonderland, and to be amused by the eccentricities on display. By the time Alice goes through the looking glass in the sequel, however, she is allowed to laugh, no fewer than ten times. Arguably, the reader is now invited to share Alice's laughter and thus have a more empathic relationship with Alice than in the preceding novel. As Kramer (2012: 289) observes: "The contagious factor of laughter is relevant to a discussion on empathy and its role in intersubjectivity as a mechanism to bring people together in shared experience".

In *Through the Looking Glass* Alice's laughter functions in part to model the reader's response to the situations she is encountering. For example, when the king is unable to mount his horse without falling over, she recommends that he acquire a wooden one with wheels. He then asks her if such a horse goes smoothly.

(5) 'Much more smoothly than a live horse,' said Alice, with a little scream of **laughter**, in spite of all she could do to prevent it.

The laughter in this episode indicates to young readers that the slapstick tumbling of the figure of authority from his horse is funny, but also that the uncontrollable scream of laughter that they might share with Alice is something that should ideally be disciplined or even prevented. In *Through the Looking Glass*, as in most of the other novels in the corpus, the controlling of laughter and the expression of proper kinds of laughter are associated with the child's conformity with the social norms of the adult world.

6. Concluding comments

In the spirit of distant reading, then, leavened with a necessary scepticism about the very categories our data produces, we offer the following comments on the significance of the findings outlined in the sections above. The categories of laughter that we have explored are based, first of all, on two extensive lexicographical works of reference, each of which took generations of intellectual labour to produce: the *OED* and *HTE*. Despite their justifiable reputation as authoritative works of reference, neither is completely consistent or entirely comprehensive. Even so, the lexicographical references offer, we argue, a marvellously rich point of departure for corpus stylistics.

Once the search items were selected, drawing on the lexical domain of *laughter* as delineated by the *HTE*, the text analyses produced a set of results that required interpretation. No interpretation can simply be 'read off' the tables and graphs we have produced (cf. Fish 1980); each individual reading demands some previous knowledge of what the items might mean and assumptions about how and why they might be used in the texts. Even so, we were surprised by some of the findings and consequently revised our understanding of the uses of laughter in the texts. We were puzzled by the total absence of any lexicalisations of laughter in *Alice's Adventures in Wonderland*; we also assumed that there would be higher numbers of words expressing specific types of laughter across the novels analysed. We expected more *sneering, cackling* and *giggling* than we found. What we did discover was evidence of the ways in which different novels draw upon the lexical domain of laughter in their narratives, and how the expressions contribute to readers' sense of community and morality. Our findings ended up affirming our initial hypothesis that laughter in children's literature is the echo-chamber of the soul.

REFERENCES

Sources

Alcott, L.M.
1868 *Little Women*. https://www.gutenberg.org/ebooks/514, accessed September 2020

Barrie, J.M.
1911 *Peter Pan and Wendy*. https://www.gutenberg.org/ebooks/26654, accessed September 2020

Baum, L.F.
1900 *The Wonderful Wizard of Oz*. https://www.gutenberg.org/ebooks/55, accessed September 2020

Books in Children's Literature
https://www.gutenberg.org/ebooks/bookshelf/20, accessed September 2020

Burnett, F.H.
1911 *The Secret Garden*. https://www.gutenberg.org/ebooks/113, accessed September 2020

Carroll, L.
1865 *Alice's Adventures in Wonderland*. https://www.gutenberg.org/ebooks/928, accessed September 2020
1871 *Through the Looking Glass*. https://www.gutenberg.org/ebooks/12, accessed September 2020

Coolidge, S.
1872 *What Katy Did*. https://www.gutenberg.org/ebooks/8994, accessed September 2020

Grahame, K.
1908 *The Wind in the Willows*. https://www.gutenberg.org/ebooks/289, accessed September 2020

Historical Thesaurus of English
http://historicalthesaurus.arts.gla.ac.uk, accessed September 2020

Kay, C. et al. (eds.)
2009 *Historical Thesaurus of the Oxford English Dictionary*. Oxford: Oxford University Press.

Lofting, H.
1920 *The Story of Doctor Dolittle*. https://www.gutenberg.org/ebooks/501, accessed September 2020

London, J.
1903 *The Call of the Wild*. https://www.gutenberg.org/ebooks/215, accessed September 2020

Oxford English Dictionary (OED online)
http://www.oed.com, accessed September 2020

Sewell, A.
1877 *Black Beauty*. https://www.gutenberg.org/ebooks/271, accessed September 2020

Stevenson, R.L.
1883 *Treasure Island*. https://www.gutenberg.org/ebooks/120, accessed September 2020

Wiggin, K.D.
1903 *Rebecca of Sunnybrook Farm*. https://www.gutenberg.org/ebooks/498, accessed September 2020

Special studies

Anderson, W. – E. Bramwell – C. Hough
2016 *Mapping English Metaphor through Time.* Oxford: Oxford University Press.

Anthony, L.
2019 *AntConc* (Version 3.5.8). Tokyo: Waseda University. Available from https://www.laurenceanthony.net/software, accessed September 2020

Bartsch, S.
2004 *Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence.* Tübingen: Gunter Narr Verlag.

Busse, B.
2012 "A celebration of words and ideas: The stylistic potential of the *Historical Thesaurus* of the *Oxford English Dictionary*", *Language and Literature* 21 (1), 84-92.

Carpenter, H.
1985 *Secret Gardens: A Study of the Golden Age of Children's Literature*. London: Faber & Faber.

Corbett, J.
2021 "Terminology and the evolution of linguistic prejudice: The conceptual domain of 'Irishness' in the *Historical Thesaurus of English* and the *Hansard Corpus of British Parliamentary Speeches*", *TradTerm* 37 (2), 515-537.

Darton, F.J.H.
1982 [1932] *Children's Books in England: Five Centuries of Social Life* (3rd edn.). Cambridge: Cambridge University Press.

Firth, J.R.
1957 *Papers in Linguistics, 1934-1951*. Oxford: Oxford University Press.

Fish, S.
1980 *Is There a Text in this Class? The Authority of Interpretive Communities*. Cambridge, Mass.: Harvard University Press.

Hunston, S.
2007 "Semantic prosody revisited", *International Journal of Corpus Linguistics* 12 (2), 249-268.

Kramer, C.A.
2012 "As if: Connecting phenomenology, mirror neurons, empathy, and laughter", *PhaenEx* 7 (1), 275-308.

Moretti, F.
2005 *Graphs Maps Trees: Abstract Models for Literary History.* London: Verso.

Oster, U.
2010 "Using corpus methodology for semantic and pragmatic analyses: What can corpora tell us about the linguistic expression of emotions?", *Cognitive Linguistics* 21 (4), 727-763.

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

Sorby, A.
2011 "Golden Age." In: P. Nel – L. Paul (eds.) *Keywords for Children's Literature.* New York: New York University Press, 96-99.

Stewart, D.
2010 *Semantic Prosody: A Critical Evaluation.* London; New York: Routledge.

APPENDIX 1

Laugh (intransitive verb) in a specific manner (*HTE* 02.04.10.11.01vi)

Lexical item	Currency	Meaning	*HTE* code
roar	1815-	Laugh loudly/ coarsely	02.04.10.11.01.01vi
snort	1825-		
ha-ha	1320-		
die with/ of laughing	1596-	Laugh convulsively, immoderately	02.04.10.11.01.02vi
giggle	1609-	Giggle	02.04.10.11.01.03vi
titter	1619-		
sneer	1683-	Laugh foolishly	02.04.10.11.01.05vi
whinny	1825-	Laugh in the manner of a horse	02.04.10.11.01.08vi
snicker	1694-	Snigger	02.04.10.11.01.09vi
snigger	1706-		
cackle	1712-		
chuckle	1803-	Chuckle	02.04.10.11.01.10vi
chortle	1871-		

APPENDIX 2

Types of laughter (HTE code 02.04.10.11.01n)

Lexical item	Currency	Meaning	*HTE* code
roar	1778-	instance/outburst of loud/coarse/ immoderate laughter	02.04.10.11.01.01.01n
ha ha	1806-		
convulsion(s)	1735-	outburst of vehement/ convulsive/ wild laughter	02.04.10.11.01.02.01n
giggling	a1510 +1786-	giggling/tittering	02.04.10.11.01.01.05n
tittering	1657-		
giggle	a1677-		
titter	1728-		
snickering	1775-	sniggering	02.04.10.11.01.01.07n
sniggering	1775-		
snigger	1823-	instance of sniggering	02.04.10.11.01.01.07.01n
chuckling	1820-	chuckling	02.04.10.11.01.01.08n
chuckle	1837-		
chuckle	a1754	instance of chuckling	02.04.10.11.01.01.08.01n
cackle	1856-		

APPENDIX 3

Raw and normalised frequencies of search items in the novels

Text Code	AW		LG		LW		WKD		BB		TI		WWO		CW		RSF		WW		PPW		SG		DD	
Raw/Normalised frequencies	R	N	R	N	R	N	R	N	R	N	R	N	R	N	R	N	R	N	R	N	R	N	R	N	R	N
laugh*	0	0	17	555.2	239	1250	44	860.6	25	410.9	17	242.0	22	551.5	15	463.5	28	368	46	757.1	18	373.6	107	1286.6	11	399
roar*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	16.5	0	0	0	0	0	0
snort*	0	0	0	0	0	0	0	0	1	16.4	0	0	0	0	1	30.9	0	0	1	16.5	0	0	0	0	0	0
ha ha*	0	0	0	0	4	20.9	0	0	1	16.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	108.8
die* with/of laughing	0	0	0	0	1	5.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
convuls*	0	0	0	0	7	36.6	3	58.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
giggl*	0	0	0	0	4	20.9	9	176.0	0	0	0	0	0	0	0	0	1	13.1	6	98.8	0	0	5	60.1	1	36.3
titter*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	13.1	0	0	0	0	0	0	0	0
sneer*	0	0	0	0	0	0	0	0	0	0	4	56.9	0	0	2	61.8	0	0	0	0	4	83	1	12	1	36.3
whinn*	0	0	0	0	0	0	0	0	1	16.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
snicker*	0	0	0	0	0	0	1	19.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
snigger*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	65.8	0	0	0	0	0	0
cackl*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	13.1	0	0	0	0	0	0	0	0
chuckl*	0	0	0	0	3	15.7	1	19.6	1	16.4	2	28.5	0	0	1	30.9	1	13.1	5	82.3	1	20.8	12	144.3	0	0
chortl*	0	0	1	32.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Raw frequency = number of occurrences in the text
Normalised frequency = number of occurrences per 1,000,000 words

Address: John Corbett, BNU-HKBU United International College, English Language & Literature Studies, Zhuhai, Guangdong, China.
ORCID code: https://orcid.org/0000-0002-5805-1607

Address: Li Li, Macao Polytechnic University, Faculty of Languages and Translation, Macao, Macao SAR.
ORCID code: https://orcid.org/0000-0003-3652-3498

English or Maltese? Language use among university students on social media platforms

Mario Vassallo and Lydia Sciriha

University of Malta

ABSTRACT

Malta's Constitution declares both Maltese, the indigenous language, and English as the country's official languages. Maltese is also the national language and since 2002 it was accorded official status in the European Union. Maltese is therefore given more importance in Malta, a miniscule island with a population of slightly more than half a million people.

This study reports on the findings of a scientifically representative study among 500 University of Malta students on their language use when using social media platforms. It provides data on the actual languages used in messages sent by the students themselves. This paper examines the different contexts in which English and Maltese are used on the social media platforms. It compares how the participants spontaneously use either language in different social media forms of communication. The study concludes that rather than a process of displacement of Maltese, what is happening is differential usage through which Maltese is predominant in informal settings, while English is mainly used in more formal settings.

Keywords: Maltese, social media platform, official language, domain, frequency of use, language proficiency.

1. The context

The Maltese language is one of the lesser used languages of the world: it is spoken in Malta by its inhabitants and, much less so, by the Maltese diaspora. Traditionally, Maltese served as the identity carrier for the inhabitants of the Maltese archipelago during centuries of foreign rule. It

DOI:10.25951/4822

served as an important buffer against the influences of the foreigner who occupied the islands because of their geographically strategic significance. Maltese is a Semitic language, but over the centuries it has accumulated a significant amount of Romance and Anglo-Saxon accretions. Historically too, it was often derided by many who did not appreciate its intrinsic value, and by others who would have preferred that, politically, the island should not be on its own in the middle of the Mediterranean Sea known for its inter-tribal animosities, but rather belong to a larger nation, often on the pretext that such a small population was not sustainable. Maltese was often instrumentalised politically to promote the interests of the British colonisers to displace the dominant position of Italian. In this regard Malta has a well-documented 'language question' (Hull 1993).

Despite this, Maltese did survive. It began to acquire added prominence when it started to be written and disseminated through print. Maltese vacillated from being considered useful only for 'use in the kitchen', as it was frequently derided before the two world wars, to being recognised as one of the official languages of the European Union when Malta was accepted as a full member of the Union. Maltese pragmatists naturally recognised that Maltese could not be their only language if they wanted to be able to communicate, without losing their identity however, with the outside world. Because of this, Italian had developed in Malta, in parallel to its development in Italy, and was extensively used by the 'literati' and the Church. With the massive effect of the second world war when the Maltese fought a war which was not theirs, English started to be more widely accepted, and today constitutes Malta's other 'official' language, second to Maltese which is also Malta's 'national' language (Vassallo 1979).

With globalisation, the advent of universal education and the massive spread of the new media of communication and of mass tourism, the use of English in Malta rapidly increased as it did in other societies. This naturally resulted in the loss of space for Maltese, and both quantitative and qualitative linguistic research (e.g. Boffa 2010; Brincat 2005; Caruana 2006; Farrugia 2019; Sciriha 2016; Sciriha – Vassallo 2001, 2006; Vassallo – Sciriha 2020) on the recent experience of the language started to point to a resultant meltdown of the language. The reasons put forward in this research were based on both external and internal factors: on the one hand the external ones were related to the need for the Maltese to be active as citizens of the world; on the other hand, internal factors were based on the somewhat cavalier use of the language by the inhabitants of the island, especially in communication with children, and the constant use of code switching entertained by speakers of all social groups. The question therefore arises as to whether English will eventually eclipse Maltese.

This paper addresses a number of issues. It first seeks to establish the self-perceived language proficiency among the Maltese on social media platforms. On the basis of the data collected, it then seeks to establish whether the media are instrumental in pushing Maltese into disuse, or whether their choice of language actually reflects a much wider preference for differential use of languages in specific contexts, as has been documented in other bilingual contexts. According to Fishman (1965) language choice is not random but there is a pattern in such a choice which is governed by what he calls 'domains'. These are institutional contexts in which one language is likely to occur more than the other. Some domains, such as the family domain, are less formal than others and there is differential preference. Other studies (Bishop – Hicks 2005; Costa – Santesteban 2004; Gonzalez-Vilabazo – López 2012) show that adult proficient bilinguals tend to allow themselves to use both languages interactively, with code-switching being the most common practice when communicating with their in-groups (Bhatt – Bolonyai 2011). These studies suggest that, whilst formal language is used whenever interlocutors are not familiar with each other, an element of 'laissez-faire' in language choice, or language 'combinations' is adopted in communicating with persons who are so intimately known to each other. The 2016 study by Jongbloed-Faber and others on the use of Frisian teenagers in social media suggests that Frisian use is expanding despite the fact that Frisian is mostly spoken and not written. The Jongbloed-Faber group explain that Frisian is the mother tongue of 54% of the 650,000 inhabitants of the province and is predominantly a spoken language. Actually, 64% of the Frisian population can speak it well, while only 12% indicate that they can write well. But their study shows that as many as 87% of this group use it to some extent as their medium of communication on social media. It is specifically this aspect of differential use of the two dominant languages in Malta that this study seeks to address.

To answer this set of questions, a quantitative study was undertaken among a representative sample of Maltese university students. It is commonly held that what goes on among this 'elite' group of citizens is likely to be the foretaste of things to come. The findings of the study will be used in this paper to document the relative use of English among this group, how they evaluate it, and what preferences they have in various social media domains.

2. Language use on social media platforms?

In everyday face-to-face conversations, bilingual speakers are always faced with an important choice. Which one of the languages in their linguistic repertoire do they select and for what reason? Very often they converge towards the needs of the addressee (Giles et al. 1977) and only rarely do they consciously decide not to accommodate the addressee's needs.

Studies on language use on social media platforms are by no means as prolific when compared to those which focus on bilinguals' language use in different domains (e.g. family, transactions, education, church). Only recently has the use of English on social media platforms been studied by researchers such as Kelly-Holmes (2019), while Jongbloed-Faber et al. (2016) and Jongbloed-Faber (2021) investigated the use of Frisian among teenagers in the province of Fryslân in the Netherlands.

The present study investigates the use of languages among students at the University of Malta in respect of three social media platforms: WhatsApp, Facebook and Twitter. Of these three platforms, respondents use Facebook and WhatsApp profusely, whilst Twitter is not so popular. In the case of bilinguals, one language tends to be preferred on the basis of whether the message is private or public. Successive surveys have been conducted in Malta over the last two decades (Sciriha 1998, 2001, 2018; Sciriha – Vassallo 2001, 2006) to examine the use of the official languages in different domains. To date, no study has been conducted on language use on social media, despite their proliferation and accessibility even to people who are geographically far away from each other.

3. Methodology

A scientifically representative survey was conducted among 500 University of Malta students following courses in fourteen Faculties just before the Covid-19 pandemic. In-person interviews with the selected students were held on campus by a team of interviewers. The instrument used to collect the data was a structured questionnaire in which, besides the demographic data, respondents were asked questions pertaining to their mother tongue, their parents' occupation and the faculty they belonged to. Other questions focused on (i) their self-rated proficiency levels in the spoken and written skills and (ii) the frequency of use of these skills. The main focus of the study was the students' language use in either English and/or Maltese on social media platforms, more specifically their usage on different platforms such as Facebook, WhatsApp and Twitter.

Table 1 gives a sample profile by gender and faculty. More females (N = 295) were interviewed because the overall total percentage of female students at the university is higher than that of male students (N = 205).

Table 1. Sample profile, by gender and faculty

Faculty	Male	Female	Total
Arts	20	41	61
Column %	9.8	13.9	12.2
Built Environment	12	8	20
Column %	5.9	2.7	4.0
Dental Surgery	2	6	8
Column %	1.0	2.0	1.6
Education	4	19	23
Column %	2.0	6.4	4.6
Engineering	17	5	22
Column %	8.3	1.7	4.4
FEMA (Management & Accountancy)	43	44	87
Column %	21.0	14.9	17.4
Health Sciences	18	53	71
Column %	8.8	18.0	14.2
Information Technology	13	3	16
Column %	6.3	1.0	3.2
Laws	15	26	41
Column %	7.3	8.8	8.2
Media & Knowledge Science	7	9	16
Column %	3.4	3.1	3.2
Medicine & Surgery	27	33	60
Column %	13.2	11.2	12.0
Science	12	10	22
Column %	5.9	3.4	4.4
Social Wellbeing	13	37	50
Column %	6.3	12.5	10.0
Theology	2	1	3
Column %	1.0	0.3	0.6
Total	**205**	**295**	**500**

The female presence is higher in the Faculties of Arts (N = 41 vs. 20 males), Education (N = 19 vs. 4 males), Wellbeing (N = 37 vs 13 males), Health Sciences (53 vs. 18 males), Medicine and Surgery (N = 33 vs. 27 males), and Dental and Science (N = 6 vs. 2 males). The number of female students is lower in other faculties, particularly so in Engineering (N = 5 vs. 17 males) and Information Technology (N = 3 vs. 13), to mention two. Random stratified sampling was used to ensure that the base reflected the total full-time student population at the University of Malta.

4. Mother tongue and language preference

To put the study in perspective, the participants were asked about what they considered their mother tongue. This was defined as 'the language learnt from parents/guardians as a child'. They were subsequently also asked what language they actually preferred to speak. Table 2 presents the findings about the students' reporting of what their mother tongue is, broken down by the socio-economic group of their household.

The figures in Table 2 clearly indicate that the majority of the students (73.0%) were brought up in families in which both parents spoke Maltese. Families in which both parents spoke only English add up to only 8.2% of the total sample. Some 16.4% of the sample originated from a Maltese and English bilingual household whilst the rest (2.4% in all) hailed from families with other language combinations. It is worthwhile noting that of the entire sample only 1% had a background which did not include any Maltese or English.

When the participants were in turn asked what language they prefer to speak, as many as 50.8% of all the respondents stated that they prefer to speak in Maltese, in contrast to 21% who claimed that they prefer to speak in English. Another 28.2% do not have any specific preference, thus indicating that they feel that they are balanced bilinguals. The full details are presented in Table 3.

With a *p-value* of 0.000, the relationship between preferred language and the respondents' household socio-economic category is significant. In this respect, it is obvious from the table that the lower the socio-economic category, the higher the preference for Maltese as the medium of 'spoken' communication: the number of those who consider Maltese as their mother language within the lowest socio-economic group, the DE group, amounts to 72.7%, in contrast to only 3.6% of the same group who consider English as their mother tongue. Interestingly, the percentages of

Table 2. Mother tongue, by household socio-economic group

	TOTAL	Household Socio-Economic Category			
		AB	C1	C2	DE
Maltese	365	118	130	63	54
Row %	100.0	32.3	35.6	17.3	14.8
Column %	73.0	58.7	78.3	80.8	98.2
English	41	25	12	4	0
Row %	100.0	61.0	29.3	9.8	0.0
Column %	8.2	12.4	7.2	5.1	0.0
Maltese & English	82	50	21	10	1
Row %	100.0	61.0	25.6	12.2	1.2
Column %	16.4	24.9	12.7	12.8	1.8
English & another language	3	2	0	1	0
Row %	100.0	66.7	0.0	33.3	0.0
Column %	0.6	1.0	0.0	1.3	0.0
Maltese & another language	4	2	2	0	0
Row %	100.0	50.0	50.0	0.0	0.0
Column %	0.8	1.0	1.2	0.0	0.0
Other languages, neither Maltese nor English	5	4	1	0	0
Row %	100.0	80.0	20.0	0.0	0.0
Column %	1.0	2.0	0.6	0.0	0.0
Total	**500**	**201**	**166**	**78**	**55**

Table 3. Preferred language for spoken communication, by household socio-economic group

	TOTAL	Household Socio-Economic Category			
		AB	C1	C2	DE
Maltese	254	65	90	59	40
Row %	100.0	25.6	35.4	23.2	15.7
Column %	50.8	32.3	54.2	75.6	72.7
English	105	71	25	7	2
Row %	100.0	67.6	23.8	6.7	1.9
Column %	21.0	35.3	15.1	9.0	3.6
Either Maltese or English: No Difference	141	65	51	12	13
Row %	100.0	46.1	36.2	8.5	9.2
Column %	28.2	32.3	30.7	15.4	23.6
Total	**500**	**201**	**166**	**78**	**55**

the highest socio-economic group, the AB respondents who prefer to speak in English (at 35.3%), is not very different from those representing speakers who prefer to use Maltese (at 32.3%) as the medium for their spoken communication.

5. Language proficiency in and frequency of use of English and Maltese

Respondents were also asked to self-evaluate their spoken and written proficiency levels in both official languages. Though this exercise is fraught with difficulties since persons usually tend to inflate their proficiency levels in languages, yet it gives researchers an idea of the respondents' proficiency levels in the two languages. Moreover, this exercise provided students with the opportunity to reflect on their competencies in the two official languages. As evident in Table 4, a high 77.8% of the university students reported speaking Maltese 'very well', while 13.2% speak Maltese 'well'. Only 2.4% declared that they spoke Maltese 'with some difficulty'. As regards their writing skills in the national language, their levels of proficiency are lower than their speaking ones. Still, slightly more than the majority of the students (58.8%) said they write Maltese at the highest level of proficiency ('very well') and 24.4% evaluated their written Maltese at a lower level (well: 24.4%). Interestingly, in respect of English, their writing skills at the highest level surpass those in Maltese (English 'very well': 69.4% vs. Maltese: 58.8%). Moreover, while only 2.2% of the students said that they write 'with difficulty' in English, this figure was higher for Maltese (5.0%).

Table 4. Proficiency in spoken and written Maltese and English

Proficiency Levels	MALTESE		ENGLISH	
	Speaking	Writing	Speaking	Writing
	%	%	%	%
Very well	77.8	58.8	69.0	69.4
Well	13.2	24.4	23.6	21.8
Reasonably well	5.6	10.8	6.0	5.8
With difficulty	2.4	5.0	0.4	2.2
None	1.0	1.0	1.0	0.8
Total	**100.0**	**100.0**	**100.0**	**100.0**

In order to reveal what these values actually mean, a 100-Point Proficiency Index was constructed, and is presented in Table 5. The Index was constructed through a weighting system that differentiates the values obtained through the Likert Scale summarised in Table 4. The Index shows that in a range of –100 to +100, respondents estimate their proficiency in speaking Maltese to exceed their proficiency in speaking English, at 91.10 and 89.80 points respectively. Both values are very high. What is quite interesting is that the two Indices are so close to each other, which clearly suggests that altogether Maltese tertiary students consider themselves to be balanced bilinguals in the spoken domain.

Table 5. 100-Point Language Proficiency Index

	PROFICIENCY INDEX
Maltese Speaking	91.10
Maltese Writing	83.75
English Speaking	89.80
English Writing	89.20

The Index figures are slightly lower in respect of writing, and, not surprisingly, the Index for writing in English is higher than the Index for writing in Maltese, at 89.20 and 83.75 points respectively. The reason for this is that although Maltese is phonetically written, the existence of the two typically Semitic unsounded consonants (*h* and *għ*) present significant orthographic difficulties.

In addition to their evaluation of proficiency, respondents were also asked about the extent of their use of English and Maltese. The data in Table 6 show that spoken Maltese is more frequent than spoken English ('All the time': Spoken Maltese 78% vs. Spoken English: 62%) among the respondents. However, the frequency of writing in English is much higher than it is in Maltese. English clearly outstrips Maltese in so far as frequency of writing is concerned. In fact, 73.2% of the respondents claimed to write in English 'all the time' when compared to 51% of those who use Maltese in writing. Moreover, 7.6% said that they 'never' write in Maltese. Only 0.2% of the respondents claimed never to write in English.

Once more the findings summarised in Table 6 were computed into another 100-Point Index, and they are presented in Table 7. On this Index, the values for Maltese and English speaking are respectively 91.60 and 86.95 points. Writing in English, however, exceeds writing in Maltese by 14.80 points, which is very significant. This shows that in written communication English is extensively preferred to Maltese. In view of the fact that Social Media practically rely on the written form of language use, this already points to important considerations in answer to the questions set for this paper. The data suggest that, rather than driving language shift, social media preference conforms to a wider pattern of language use. The data suggest that Maltese is preferred for private, domestic and local use while English is preferred for public and international use. Social media has an immediacy of interaction more typical of speech than of writing and in this context. As such, Maltese is preferred when social media is more 'speech-like', while English is preferred when social media is more 'writing-like'. This points to an interesting functional differentiation process of the two languages in social media preferences.

Table 6. Frequency of use in spoken and written Maltese and English

Frequency	MALTESE		ENGLISH	
	Speaking	Writing	Speaking	Writing
	%	%	%	%
All the time	78.0	51.0	62.0	73.2
Often	13.6	15.0	26.6	19.2
Now & then	6.2	23.8	9.4	6.4
Never	1.2	7.6	1.2	0.2
NA	1.0	2.6	0.8	1.0
Total	**100.0**	**100.0**	**100.0**	**100.0**

Table 7. 100-Point Language Usage Index

	USAGE INDEX
Maltese Speaking	91.60
Maltese Writing	76.05
English Speaking	86.95
English Writing	90.85

6. Language use on social media platforms

The primary objective of this study was to investigate the use of the two official languages on social media platforms and whether when using these platforms, respondents employ more English than Maltese. It sought to discover which one of the two languages is more prevalent on three popular platforms: WhatsApp, Facebook and Twitter. For this reason, separate questions were asked to collect hard data on whether there is a difference in the use of languages depending on the social media platforms used.

6.1 Group and private WhatsApp messages

WhatsApp allows the user to send both group messages and private messages. This distinction is important in view of the fact that a private message is sent to one addressee who, typically, is well known to the sender of the message, whereas a group message is sent to several addressees who might not all know Maltese but would be able to understand messages in English. As such, the data presented in Table 8 reveal the extent to which this is important. Whereas 41.4% of the respondents reported sending private messages in Maltese 'all the time', a lower percentage is registered in respect of English (30.4%). In contrast, with regard to Group messages, the total percentage of 67.6% regarding the use of English in two frequency levels of 'all the time' (32.2%) and 'often' (35.4%) is identical to that of Maltese language use in these two frequencies of use. This is so even though the use of Maltese is higher when it is used 'all the time' (39.8%) and lower when it is used 'often' (27.8%). In contrast, 18.6% of the respondents claimed that they 'never' send group messages in Maltese, while only 12.2% do so in English.

Table 8. Maltese and English use on group and private WhatsApp messages

Frequency	MALTESE		ENGLISH	
	Group WhatsApp	Private WhatsApp	Group WhatsApp	Private WhatsApp
	%	%	%	%
All the time	39.8	41.4	32.2	30.4
Often	27.8	25.8	35.4	33.4
Now & then	13.8	14.6	20.2	22.0
Never	18.6	18.2	12.2	14.2
Total	**100.0**	**100.0**	**100.0**	**100.0**

What these figures suggest, therefore, is that when the frequencies for 'all the time' and 'often' are taken together, Maltese occupies slightly more space on Whatsapp in respect of private messaging but occupies an equal space with English in respect of group messaging.

6.2 Facebook

One of the most popular social media platforms is Facebook. In fact, out of 500 respondents only 3 said that they do not have their own Facebook page. A high 87% said that they check their Facebook accounts 'all the time' (26.0%) or 'as often as I can' (61%).

Like WhatsApp, Facebook includes private messages and status updates which are public. In view of this distinction, respondents in the survey were asked to cite the language they use when updating their Facebook status and also when sending private messages on this platform. The findings are presented in Table 9.

English is the language that is used 'all the time' for status updates by 30.6% when compared to a mere 6.4% who use Maltese. Conversely, while 19.2% of the participants 'never' use English for status updates, a much higher percentage (45.4%) said that they 'never' use Maltese.

The situation regarding language use changes in private messages. A high 50.4% of the participants claimed to use Maltese 'all the time', when compared to a lower 36.0% who claimed to use English with the same high frequency. On the other side of the frequency spectrum, there is really not much difference between those who claimed 'never' to use either Maltese (5.8%) or English (4.0%) in private messages.

As such, Maltese is preferred in inter-personal communication on the Facebook platform, but English is the preferred medium for the propagation of one's status as reflected in the language used for regular updates.

Table 9. Maltese and English use on status updates and private Facebook messages

Frequency	MALTESE		ENGLISH	
	Facebook Status update	Facebook Private Messages	Facebook Status update	Facebook Private Messages
	%	%	%	%
All the time	6.4	50.4	30.6	36.0
Often	18.8	30.0	27.4	37.4
Now & then	29.4	13.8	22.8	22.6
Never	45.4	5.8	19.2	4.0
Total	**100.0**	**100.0**	**100.0**	**100.0**

6.3 Twitter

Twitter is not as popular among the university students. In fact, a very high 73.4% of them do not even have a Twitter account. Nevertheless, those who have such an account either use it rather sparingly or never use it.

Twitter allows two different types of tweets: regular ones and tweets with @. Regular tweets are messages posted on Twitter that could contain text, photos, a GIF, and/or video. This kind of tweet appears on the sender's profile page and Home timeline. It also appears in the Home timeline of anyone who is following the sender. In contrast, tweets with @ show in the recipient's Notifications tabs, which are accessible only to them. Additionally, mentions will appear in the recipient's Home timeline view (not on their profile) if they are following the sender. This group of tweets is seen by anyone on Twitter who is following the sender in their Home timeline. This makes tweets with @ somewhat more private.

Table 10 gives a breakdown of the findings of respondents who send either regular tweets or tweets starting with @. The results show that in respect of regular tweets which are sent 'all the time', English is used significantly more (30.6%) than Maltese (1.6%). Moreover, while 52.7% of the participants claimed that they never send a regular tweet in English, the percentage is much higher for those who 'never' send regular tweets in Maltese (83.5%).

In respect of sending tweets which start with @ with great frequency ('all the time'), again English is the preferred language: 22.6% send such tweets in English when compared to a mere 0.5% in Maltese. The percentages of English language use for 'never' sending such tweets are much lower (54.3%) when compared to Maltese (80.1%).

Table 10. Maltese and English use when using regular tweets and tweets beginning with @

Frequency	MALTESE		ENGLISH	
	Regular Tweet	Tweet starting with @	Regular Tweet	Tweet starting with @
	%	%	%	%
All the time	1.6	0.5	30.6	22.6
Often	3.2	2.2	4.3	10.8
Now & then	11.7	17.2	12.4	12.4
Never	83.5	80.1	52.7	54.3
Total	**100.0**	**100.0**	**100.0**	**100.0**

In respect of the use of Twitter, the pattern does not appear to follow that used in respect of Facebook: English is the language most often used for both kinds of tweets, whether they are the regular ones which are more public, or the more private ones that start with @. On this particular platform, English is more dominant than Maltese among the participants in this study.

7. Language ranking

The foregoing set of data is a vivid expression of the language ranking Maltese university students use in their daily lives. This is done unconsciously and unobtrusively but is very real in its consequences. To test the consistency between conscious and unconscious language ranking processes, participants in this study were specifically asked to rank seven languages according to two different factors, namely in respect of their being 'citizens of Malta', and subsequently in 'their being citizens of the world'. The findings are respectively presented in Tables 11 and 12.

Table 11. Ranking of seven languages in terms of perceived importance as Maltese nationals living in Malta among UOM students

	1st	2nd	3rd	4th	5th	6th	7th
	%	%	%	%	%	%	%
Maltese	72.4	25.0	0.2	0.4	1.0	0.4	0.6
English	32.8	65.4	1.2	0.4	–	–	0.2
Italian	0.8	4.4	79.4	10.0	3.8	1.4	0.2
French	0.0	1.2	10.0	42.2	30.2	12.2	4.2
German	0.0	0.4	3.6	11.2	26.6	33.8	24.4
Spanish	0.8	0.2	2.2	13.2	23.0	35.2	25.4
Arabic	0.0	1.0	3.8	20.4	14.4	15.2	45.2

Table 12. Ranking of seven languages in terms of perceived importance as citizens of the global society among UOM students

	1st	2nd	3rd	4th	5th	6th	7th
	%	%	%	%	%	%	%
English	96.2	3.2	0.6	0.0	0.0	0.0	0.0
French	1.6	41.0	30.8	15.2	7.0	3.4	1.0
Italian	0.0	19.0	23.4	23.6	19.0	14.0	1.0
German	0.6	7.4	19.6	28.6	28.4	13.4	2.0
Arabic	0.8	7.6	9.0	11.2	12.6	37.6	21.2
Spanish	0.8	16.0	17.4	20.6	29.2	12.4	3.6
Maltese	1.0	3.8	2.4	2.4	3.8	16.4	70.2

Table 11 and Table 12 are interesting in the way they differ from each other: in respect of being a citizen of Malta, Maltese is ranked first (72.4%,) whilst English was ranked first by 32.8% of participants. At the same time, English was ranked second by 65.4% whilst Maltese was ranked second by 25%. The space allowed for other languages at these two highest levels is minimal. This contrasts very sharply with the rankings obtained when the same set of languages were ranked according to 'being a citizen of the world'. Here English dominates, with as many as 96.2% of the participants ranking it first, in sharp contrast with only 1% of those who ranked Maltese as the most important language. In fact, as many as 70.2% ranked it as the least important language. Even though Maltese is one of the official EU languages, its relevance on the international plane is considered minimal.

8. Conclusion

This study sought to map the relevance of Maltese and English in social media communication, and to identify patterns which could point to shifts in the importance of the two languages. What clearly emerges from the data collected is that both Maltese and English remain important for the Maltese, but with very different functions. Even though the Maltese are officially bilingual, different domains prompt users to use different languages. For inter-personal communication among friends, where intimacy is important, the Maltese language tends to be more commonly used. But when communication is intended to reach a wider audience, English prevails. The diversity in function is very clear, and what the consciously documented language rankings state, was clearly echoed in the use of the social media platforms. This is not necessarily true of every person involved in this study, but the pattern is clear. Maltese has its importance, which is duly acknowledged both consciously and unconsciously, but this language is relegated to practical insignificance when a medium is required to project oneself to a wider audience, or to communicate internationally.

What does the future hold? It is difficult to forecast what will happen. In a world in which atomisation is becoming increasingly more widespread (Habermas 1989), the individual's personality tends to be lost in a plethora of different identities depending on the multiplicity of transient roles which modernity has brought about. As a result, the search for a context in which the individuality of a person is recognised and celebrated, becomes very important, and the private sphere tends to become more appreciated and tenderly safeguarded (Luckmann 1967). And for this purpose, languages that reflect this privacy, and protect the individual from the intrusions of omni-present outside influences, might become more relevant and much more sought after and appreciated than at present.

The phenomenon of giving more value to privacy, as reflected in social domains like the family and in religious practices, might also affect choices about which language one uses to express intimacy and self-expression. This study clearly points in this direction: Maltese is extremely important for many Maltese university students, but despite the difficulties they encounter in writing it considering its seemingly 'problematic' orthographic rules, it is still manifestly used more than English, albeit English is equally known and easier to write, as their preferred medium for communications with their peers on all the three social media platforms studied.

Maltese university students use the same differential mechanism that is used by the Friesland teenagers in the Jongbloed-Faber (2016) group study, even though not precisely in the same way. Many Friesland teenagers simply fall back on Frisian, which they speak but generally do not write, on their social media. Maltese university students also differentiate between languages. Both Maltese and English are written and spoken and Maltese students are competent in both. But they unconsciously tend to use Maltese, traditionally the carrier of national identity, when they are interacting with their in-group, when interactions are more 'speech-like', and they do not have to bother much about correct orthography. In contrast, they use English to signify social distance when, even on the same social media, they are using a 'written-like' mode of communication.

REFERENCES

Bhatt, R.M. – A. Bolonyai

2011 "Code-switching and the optimal grammar of bilingual language use", *Bilingualism: Language and Cognition* 14 (4), 522-546.

Bishop, M. – S. Hicks

2005 "Orange eyes: Bimodal bilingualism in hearing adults from deaf families", *Sign Language Studies* 5, 188-230.

Boffa, L.

2010 *The Discourse of Sales Interactions: A Qualitative Study. BA dissertation, University of Malta.*

Brincat, J.

2005 "Maltese – an unusual formula", http://macmillandictionaries.com/MED-Magazine/February2005/27-LI-Maltese.htm, accessed December 2021

Caruana, G.

2006 *Campus Talk: Analysing the Choice of Language and Code-Switching in the Casual speech of Maltese University Students. BA dissertation, University of Malta.*

Costa, A. – M. Santesteban

2004 "Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners", *Journal of Memory and Lang*uage 50, 491-511.

Farrugia, C.

2019 "Maltese and English 'crowded out of the system'", https://timesofmalta.com/articles/view/maltese-and-english-languages-crowded-out-of-the-system.723334%2023%20July%202019, accessed December 2021

Fishman, J. A.

1965 "Who speaks what language to whom and when?", *La Linguistique* 2 (1), 67-88.

Giles, H. – R.Y. Bourhis – D.M. Taylor

1977 "Towards a theory of language in intergroup relations". In: H. Giles (ed.) Language, *Ethnicity and Intergroup Relations.* London: Academic Press, 307-348.

Gonzalez-Vilabazo, K. – L. López

2012 "Little v and parametric variation", *Natural Language and Linguistic Theory* 30, 33-77.

Habermas, J.

1989 *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Translated by T. Burger and F. Lawrence. Cambridge, MA: The MIT Press.

Hull, G.

1993 *The Malta Language Question: A Case Study in Cultural Imperialism.* Valletta: Said International.

Jongbloed-Faber, L.

2021 *Frisian On Social Media: The Vitality of Minority Languages in a Multilingual Online World*. PhD dissertation, Maastricht University.

Jongbloed-Faber, L. – H. Van de Velde – C. Van der Meer – E.L. Klinkenberg

2016 "Language use of Frisian bilingual teenagers on social media", *Treballs de Sociolingüística Catalana* 26, 27-54.

Kelly-Holmes, H.

2019 "Multilingualism and technology: A review of developments in digital communication from monolingualism to idiolingualism", *Annual Review of Applied Linguistics* 39, 24-39.

Luckmann, T.

1967 *The Invisible Religion: The Problem of Modern Society*. New York: The Macmillan Co.

Sciriha, L.

1998 "Is-Soċjolingwistika". In: K. Borg (ed.) *Lingwa u Lingwistika*. Valletta: Klabb Kotba Maltin, 192-211.

2001 "Trilingualism in Malta: Social and educational perspectives", *International Journal of Bilingual Education and Bilingualism* 4 (1), 23-27.

2016 "Is English or Maltese the *de facto l*anguage in postcolonial bilingual Malta?". In: S. Borg Barthet – I. Callus (eds.) *Crosscurrents in Postcolonial Memory and Literature: A Festschrift for Daniel Massa*. Msida: Malta University Press, 139-160.

2018 "To what extent do the types of schools shape respondents' perceived usefulness and use of English in Bilingual Malta?", *Annales Universitatis Paedagogicae Cracoviensis Studia Anglica 8,* 7-20.

Sciriha, L. – M. Vassallo

2001 *Malta – a Linguistic Landscape.* Malta: Socrates.

2006 *Living Languages in Malta*. Malta: Printit.

Vassallo, M.

1979 *From Lordship to Stewardship. Religion and Social Change in Malta*. The Hague: Mouton Publishers.

Vassallo, M. – Sciriha, L.

2020 "The meltdown of Maltese – a language perspective". In: L. Sciriha (ed.) *Comparative Studies in Bilingualism and Bilingual Education.* Newcastle upon Tyne: Cambridge Scholars Publishing, 27-50.

Address: Mario Vassallo, Department of Sociology, Faculty of Arts, University of Malta, Msida MSD2080, Malta.
ORCID code: https://orcid.org/0000-0002-6968-3570

Address: Lydia Sciriha, Department of English, Faculty of Arts, University of Malta, Msida MSD2080, Malta.
ORCID code: https://orcid.org/0000-0001-7252-7395

"Santa mozzarella!": The construction of Italianness in *Luca* (Disney and Pixar, 2021)

Davide Passa

Sapienza University of Rome/University of Silesia

ABSTRACT

Eighty-one years after *Pinocchio* (1940), Walt Disney and Pixar are back with a new animated film set entirely in Italy, *Luca* (2021). It is a coming-of-age story based on a deep friendship between two sea monster boys and a human girl from Portorosso, an imaginary sea coastal town in the Cinque Terre (Liguria), in the nostalgic mid-1950s. This study intends to investigate the construction of Italianness in the film on the visual and acoustic levels. First, this article will briefly examine the visual representation of Italian people, objects and traditions that contribute to the overall construction of fictional Italianness in the film. Then, the fictional language used to characterise the inhabitants of Portorosso to distinguish them from the sea monsters will be examined in more detail; this will be done by analysing code-switching instances, where Italianisms will be included in four different categories. Unlike *Pinocchio*, *Luca*'s producers have created an artificial code that is of particular interest to researchers in the field of Sociolinguistics and Audiovisual Studies. This study will mainly focus on the construction of identity and Kozloff's functions of film dialogues. In the final sections of the article, which will analyse code-switching, Brown and Levinson's impoliteness theory will also be addressed.

Keywords: sociolinguistics, English linguistics, film studies, characterisation, fictional language.

1. Introduction

Despite the repetition of common stereotypes typical of audiovisual products (Lippi-Green 2012), *Luca* can be considered a tribute to Italian culture. It is an American fantasy film which was produced by Pixar Animation Studios and distributed by Walt Disney Studios Motion Pictures in June 2021. It is the

DOI: 10.25951/4823

second full-length Pixar film after *Soul* that, due to Covid-19, has not been released in theatres but exclusively on Disney+. The film is directed by the Genoese Enrico Casarosa, written by Jesse Andrews and Mike Jones, and produced by Andrea Warren. It is set on the Italian Riviera, in the fictional coastal village of Portorosso (the name is a blend of those of the Italian villages of Portovenere and Monterosso), inspired by the Cinque Terre in Liguria, in northwest Italy. The setting and soundtrack are inspired by Italian society of the mid-1950s, "a golden age that feels timeless" (Nemiroff 2021), giving the film a touch of Italian nostalgia that is often stereotypically used in audiovisual products. The film centres on Luca, an adorable 13-year-old sea monster boy who lives with his species under the water's surface. The sea monsters have the ability to assume human form on land, unless water touches their body, which causes them to revert to sea monster form. In a reverse perspective, however, the audience sympathises with the sea creatures and sees humans as a threat to them. The people of Portorosso, especially the fishermen sailing distant waters, occasionally spot them and put the sea monsters' lives at risk. The barriers between humans and sea monsters are broken down at the end of the film when the two species finally coexist in the town of Portorosso and several minor characters – especially lovely old Italian women – reveal their true nature as sea monsters. In the style of a coming-of-age story, Luca meets Alberto, another sea monster boy, who encourages the protagonist to venture out of the sea for the first time and shows him his ability to look human when dry, which had been kept secret by his family. The film can be interpreted on several levels, and the messages that it conveys are equally diverse – deep friendship, inclusion, exaltation of differences, love. As the producer Andrea Warren explains, "we always liked the idea that the metaphor of being a sea monster can apply to so many different things. There is a theme of openness, showing oneself and self-acceptance, as well as community acceptance" (Jefferson 2021). In an over-interpretive mood, *Luca* has been seen as the story of a homosexual relationship between Luca and Alberto, who, in order to be accepted by society, hide their true sea monster identity; others have interpreted the film as a metaphor for refugees and immigrants. The director welcomed all these interpretations, but admitted that it was all unintentional (*Style* 2021). In a nostalgic vein, however, *Luca* is primarily a tribute to Italy in its depiction of a typical 1950s-1960s Italian summer by the sea, "the summers of our youth – those formative years when you're finding yourself" (Jefferson 2021). In preparation for the film, Disney and Pixar sent some of the film's artists on a research trip to the Cinque Terre, where they took photos of the region's landscape and people. For many of them, the trip to Italy meant visiting the places their ancestors came from, and their emotional engagement is evident in the film. The portrayal of the landscape and the people is meticulous, if often clichéd.

2. Aim and methodology

This article will analyse how Italianness is constructed in *Luca*. After watching the film several times in the original language – i.e. in English – I have noted a variety of elements used to portray fictional Italian people, culture and society. Since this work aims to contribute to the existing literature on linguistic issues, it will only briefly touch on non-linguistic elements and will focus on the fictional language used to represent Italian characters. The analysis will mainly deal with:

(a) Giulia Marcovaldo, an Italian girl who loves adventure and the entire universe. She studies in Genoa, the city of great opportunities, but returns to Portorosso in the summer to live with her father – a fisherman; she helps him deliver fish to people's houses in Portorosso. She befriends Luca and Alberto and tries to help them win the Portorosso Cup.

(b) Massimo Marcovaldo, Giulia's father. Despite his size and skill with a knife, he has a soft heart, especially for his daughter.

(c) Ercole Visconti is the local bully of Portorosso, a repeat winner of the Portorosso Cup race. He has two henchmen, Ciccio and Guido, ready to do his bidding. Ercole Visconti embodies the Italian guy who loves fashion, wears expensive clothes and loafers and bullies other kids who do not wear fashionable and expensive clothes.

(d) People on the street.

This study is mainly concerned with the construction of identity – primarily through language – and Kozloff's functions of film dialogue; Brown and Levinson's impoliteness theory will be applied in the final sections of the study, where instances of code-switching are examined and Italianisms are presented in four different categories. The originality of this paper lies in the fact that it provides a detailed description of the visual and especially acoustic strategies used to construct fictional Italian characters in a 2021 Disney and Pixar film. Eighty-one years after the release of *Pinocchio* (1940), Disney and Pixar are back with another film set entirely in Italy. However, as will be discussed in the following sections, *Luca* differs from *Pinocchio* in the way it portrays Italians linguistically.

3. Fictional voices

Voice is one of the means available to directors to give their characters their own personality and identity. The quality of the voice changes depending on various sociolinguistic variables such as age, social and geographical origin, gender, and sex, but also on factors not directly related to society and culture, such as the speaker′s emotional state, health condition, or distinctive pronunciation (e.g. stuttering, sigmatism). The use of non-standard voices is of particular importance in audiovisual products, where they are often used to convey social and dialectal features of fictional speakers, and "especially in contemporary fictional dialogue, build on a network of references and allusions which are deeply embedded in a precise regional and social context" (Montini – Ranzato 2021: 2). Although the term "voice" alludes more directly to the articulatory nature of accents, it is often used to refer to both accent and dialect. Nevertheless, worth noting is Trudgill's (1994: 7) distinction between the two terms, according to which accent "simply refers to pronunciation" while dialect "has to do also with the grammatical forms that you use, as well, perhaps, as any regional vocabulary that you employ." Audiovisual products "are particularly versatile to embed and exploit the potentialities of the representation of accents and dialects: in a way which is arguably more potent than on the written page, audiences are exposed to different modes of speech, and this contributes to highlighting the relationship between standard and non-standard English" (Montini – Ranzato 2021: 4).

Sarah Kozloff (2000: 33) has presented a taxonomy in which she lists as many as nine functions of dialogue in audiovisual products. However, it is beyond the scope of this article to discuss them all, as this study focuses only on the functions that are actually used to characterise the fictional Italians in *Luca*. Additionally, Ranzato (2021: 153) maintains that dialects "can be thought of as having been devised by authors to achieve one or more of the ends listed in (Kozloff′s) taxonomy which can thus be applied to the analysis and interpretation of accented dialogue." She explains that

> non-standard accents are used especially to […] provide the necessary context for the character (*anchorage of the diegesis*): they can tell us where the story is set and the origins of the character, their regional and social milieu; and they are used to make dialogue sound more realistic (*adherence to the code of realism*) and perhaps more relevant to our current social, even political concerns. Dialects […] are used sometimes in ways that are at the opposite end of realism, even unnaturally, dissonantly, in a blatantly 'fake' way to construct an idiolect which provides very often, but not always, a comic relief (*exploitation of the resources of language*).

As will be explained in the following sections, only two of the functions listed by Kozloff are exploited in *Luca*, namely "anchorage to the diegesis" and "exploitation of the resources of language"; on the other hand, "adherence to the code of realism" is consistently disregarded.

Nevertheless, audiovisual language must be treated with caution, and any generalisation should be well considered. It departs from real language because it is non-spontaneous and pre-fabricated; it is inauthentic orality, a mere imitation of spontaneously spoken language (Pavesi et al. 2015: 7). Ferguson (1998) defines the study of fictional linguistic varieties occurring in literature as *ficto-linguistics*, and Hodson (2014: 14) explains that the designation "ficto-linguistics can be extended to include the study of language varieties in all works of fiction, including narrative poetry, film and television". Audiovisual dialogue is an "inaccurate" imitation of natural conversation that has been "scripted, written and rewritten, censored, polished, rehearsed, and performed. The actual hesitations, repetitions, digressions, grunts, interruptions, and mutterings of everyday speech have either been pruned away, or, if not, deliberately included" (Kozloff 2000: 18). Therefore, many linguistic features that recur in the use of language in the real world (e.g. hesitations, interruptions, ongoing corrections, etc.), and which are mostly invisible and taken for granted due to the improvised nature of spoken language, acquire their own meaning when they occur in audiovisual language, which is not a spoken variety *tout court*, but rather a written-to-be-spoken language; it is not un-prepared and spontaneous, therefore each linguistic and paralinguistic feature contributes to the construction of the speaker's identity. For this reason, when writing film dialogues, producers choose very carefully the linguistic features that will characterise their fictional speakers. The characters' identities are thus constructed on the basis of what they do and the way they speak.

In a poststructuralist vein, identity is not something that an individual is born with, but rather a social and cultural construction that is also based on language, as "the relationship between language and identity is rather considered as constructive" (Motschenbacher 2011: 153). It is also through language that speakers create and perform their identities, and it is also in the language that one's identities are reflected and to be found. It does not follow that the language a speaker uses results from a particular identity; rather, language is one of the ways that people have to shape their identities. Identity is not something an individual has, but something an individual does; "rather than *have* identities, people *perform* them" (McConnell-Ginet 2001: 8). The same is true for fictional people.

However, since fictional characters should be easily categorised and recognised by the audience, they are usually endowed with a reduced number of linguistic features that are reiterated in audiovisual and literary products. This is directly related to the use of stereotypes, which is a common practice in the process of media characterisation (Gross 1991: 26-27). Studies in the field of sociolinguistics have showed that the media play an important role in reinforcing linguistic stereotypes (Lippi-Green 2012), which are "uninformed and frequently culturally-biased over generalisations about subgroups that may or may not be based on a small degree of truth" (Swann et al. 2004: 298). Hall (1997: 258) claims that "stereotypes get hold of the few simple, vivid, memorable, easily grasped and widely recognized characteristics about a person, reduce everything about the person to those traits, exaggerate and simplify them, and fix them without change or development to eternity." The selective nature of stereotyping is highlighted by Ranzato and Zanotti (2018: 1), who maintain that "[r]epresentation is always the result of an act of selection of traits and features, both visual and verbal."

As will be detailed below, Italianisms usually occur when code-switching takes place, signalling that the transition from English to Italian is due either to a sudden emotional shift (e.g. fear, surprise, anger) or to a perceived intimacy, or both, mainly in the form of kinship terms (e.g. a boy addresses his father as "papà") and diminutives ("oh, piccolina mia", said by Ercole to his Vespa). As will be discussed more thoroughly, emotional shift often leads to impoliteness, where, as Brown and Levinson (1987) argue, linguistic Face Threatening Acts (FTAs) occur. People have an identity face that they seek to preserve and promote in their social relationships. Impoliteness occurs when at least one FTA is used to attack someone's face, i.e. when people's desire to be acknowledged and not to be impeded in their actions is deliberately disrespected.

4. Analysis

In this section, the film *Luca* is analysed from two main perspectives: the visual and the acoustic codes. The first includes all the visual elements that contribute to the construction of Italianness, while the second includes both the soundtrack and the fictional language spoken by the characters. Both codes are examined in order to describe in detail how the producers managed to construct Italianness in the film. It should be borne in mind that this is an American production, which affects the way Italian elements are portrayed.

4.1 Visual code

Two settings are shown in the film: Portorosso and the realm beneath the water's surface inhabited by sea monsters. The former is of particular importance to this study and is shown right from the opening scene of the film, which sets the story in an Italian seaside landscape, where the sleeping town is depicted in the moonlight as *Gelsomina*, a small boat with its *lampara*,[1] cuts through the waves. Two fishermen wear a beret, a traditional hat usually associated with men from the South – especially Sicily – but used here as a form of generalisation to characterise two men from Liguria, in north-western Italy. The location is made clear by a map written in Italian, showing Mar Ligure and Liguria. In addition, the two fishermen enjoy opera music played by an old gramophone, while nostalgically admiring the sea. The landscape is typically Italian, with marvellous sun and water, rocks, sand, Mediterranean maquis shrubland, olive trees, vines and seagulls. Posters, shop signs, books, menus displayed outside restaurants, everything containing the written language is in Italian. When Luca meets Alberto, the latter shows him a poster depicting a boy on a Vespa[2] with the motto "Vespa è libertà"[3]. Other posters show Italian food and drinks (e.g. Chinotto,[4] ice-creams) and artistic products (e.g. "Vacanze Romane",[5] "La Strada",[6] "Pinocchio", the Colosseum, Leonardo da Vinci's flying machine). The boats have Italian

[1] Fishing lamp used in the Mediterranean to attract fish.

[2] An Italian scooter brand manufactured by Piaggio. The name means wasp in Italian.

[3] "Vespa is freedom" (my translation)

[4] A traditional Italian soft drink that tastes like Coke but has a bittersweet flavour and is made from the juice of the fruit of the myrtle-leaved orange tree.

[5] A 1953 American romantic comedy film set in Rome and produced by William Wyler.

[6] A 1954 Italian film directed by Federico Fellini.

names painted on their sides (Gelsomina, Elena, Focaccia). The people in Portorosso have mainly dark hair and eyes and tanned skin, as is typical of Italians. The women wear dresses and scarves on their heads (especially old women), while the men wear berets. Some women carry baskets full of laundry on their heads; laundry is hung outside of the buildings, directly above people's heads. The buildings are old and colourful, with beautiful balconies and plants. Old men play "scopa", a traditional Italian card game, while old women comment on the passengers. Boys play football in the square, where a white Fiat 500, a red Vespa and an Ape[7] are parked. People drink espresso, eat watermelon, ice-cream, sandwiches and pasta (e.g. trenette al pesto[8]). The sea monsters visiting Portorosso enjoy Italian cuisine, as is typical for foreign tourists. Street names are marked by picturesque signs, as is typical of Italian tourist towns (but not only). Painted on the buildings are shop signs like "Bar Pittaluga", "Circolo Pescatori", "Latteria San Giorgio", "Trattoria[9] da Marina", "Bar Giotto", "Focacceria", "Pescheria", "Bar piccolo", "Alimentari[10] Rispetto". All the locals in Portorosso talk with their hands and over-gesticulate as is customary (and stereotypical?) among Italians. The price tag on a Vespa shows the amount in Lire, the old Italian currency that was replaced by the Euro in 2002.

4.2 Acoustic code

The acoustic level is the main concern of this study. Two elements contribute acoustically to the construction of Italianness in the original, English film: the soundtrack and the fictional language spoken by the characters.

4.2.1 Soundtrack

Like the visual elements analysed in the previous section, the soundtrack helps to place the story in time and create the Italian nostalgia of the mid-1950s. Foreign audiences may not be able to place the story exactly in time, as the chronological references are not explicit but rather implied by certain songs or visual elements; nevertheless, the stereotypical image foreigners have of Italy is very much in line with what Italy looked like some sixty years ago. These culture-specific references (CSRs) situate the film in the

[7] Ape is a three-wheeled light commercial vehicle, manufactured and marketed by Piaggio.

[8] It is a traditional Ligurian dish.

[9] A *trattoria* is an Italian restaurant, usually less formal than a *ristorante*;

[10] An *alimentari* is a typical Italian grocery shop.

1950s-1960s. More specifically, the posters of "Vacanze Romane" and "La Strada", both films from 1953-1954, can be seen as references successfully positioning the story in time. The songs and opera music included in the soundtrack also help to revive the nostalgic mid-1950s. However, when the scenery shifts to below the water's surface, these Italian popular songs and arias are replaced by unfamiliar celestial background sounds and songs, but using musical instruments commonly associated with Italian folk music (e.g. accordion). The film opens with "Un Bacio a Mezzanotte",[11] which immediately sets the story in place and time. In the opening scene, an old gramophone plays "O Mio Babbino Caro",[12] an aria used repeatedly in Anglophone audiovisual products to give the scene a touch of Italianness (e.g. in James Ivory's 1985 *A Room with a View*; John Huston's 1985 *Prizzi's Honor*; Steve Bendelack's 2007 *Mr. Bean's Holiday*; Olivier Dahan's 2014 *Grace of Monaco*). Giulia's father, Massimo Marcovaldo, sings "Largo al Factotum",[13] and whistles "La Donna è Mobile",[14] and the cavatina "Una Voce Poco Fa".[15] The song "Il Gatto e la Volpe"[16] is used to portray the deep friendship between Luca and Alberto. Although the song is anachronistic – as it was composed in the 1970s – it is inspired by the Cat and the Fox, two characters from the Italian novel *The Adventures of Pinocchio* (Collodi 1883), which has become one of the symbols of Italian culture in the world. "Andavo a Cento All'ora"[17] is used to portray the so-called "Italian economic miracle", an expression used to refer to the long-lasting period of strong economic growth in Italy, especially in the years 1958-1963. "Andavo a Cento All'ora" (literally, I was driving 100 km/hr) refers to speed and new means of transport such as the Vespa and Fiat 500; in the film, the song is played by Ercole Visconti's radio while he drives a loud red Vespa. Similarly, "Fatti Mandare dalla Mamma"[18] is used to refer to the typical lifestyle of the 1960s, as is "Viva la Pappa al Pomodoro",[19] which is played in a scene involving food. The credits are accompanied by "Città Vuota", an iconic song released in 1963 by the most famous Italian female singer of all time, Mina. However,

[11] A very famous Italian song by Quartetto Cetra, released in 1952.

[12] A soprano aria from the opera Gianni Schicchi by Giacomo Puccini (1918).

[13] An aria from The Barber of Seville" by Gioacchino Rossini (1775).

[14] An aria from Giuseppe Verdi's "Rigoletto" (1851).

[15] An aria from The Barber of Seville" by Gioacchino Rossini (1775).

[16] A song composed by Edoardo Bennato in 1977.

[17] A song released by Gianni Morandi in his first album, in 1963.

[18] A song released by Gianni Morandi in 1962.

[19] A song by Rita Pavone, released in 1965, when *Il giornalino di Gian Burrasca* by Vamba was adapted into a popular RAI TV-series starring Rita Pavone in the title role. "Pappa al pomodoro" is a traditional dish from Tuscany, including bread and tomato.

foreign audiences may not understand the song lyrics nor have enough Italian CSRs; nevertheless, the purpose of culture-specific elements in *Luca* is not to provide the foreign audience with content they should understand, but rather to provide formal elements that match the foreign audience's expectation of what it means to live in Italy and to be Italian. The soundtrack and visual elements (e.g. objects, food, buildings, people) do not reflect the real, contemporary Italy that is gradually losing its peculiar characteristics, as is common in many countries and especially in touristic areas. CSRs are used to create an emotional response to certain sounds and images that tend to repeat stereotypes about Italians that are used over and over again in audiovisual products. *Luca*, in fact, offers no unexpected representation of Italianness, and everything fits into the stereotypical portrayal of Italians in fiction.

4.2.2 Language

Italianness is a feature that characterises above all the inhabitants in Portorosso. Nevertheless, not only are the proper names of the inhabitants of Portorosso Italian, but also those of the sea monsters, whose surnames are often translations of fish species into Italian. The surnames of the protagonists, Paguro and Scorfano, mean "hermit crab" and "rockfish" respectively. Other sea monsters are Mr Branzino and Bianca Branzino (seabass) and Mrs Aragosta (lobster). Additional names include Caterina, Giuseppe, Enrico, Daniela, Uncle Ugo, and Mona Lisa, the last name being a reference to the painting by Leonardo da Vinci. The CSRs to the fish species in Italian will be fully understood only by Italians, who are undoubtedly those who enjoy the film the most. Nevertheless, as with the soundtrack and visual elements, the formal level of these CSRs – i.e. the exotic sound of the surnames – will help to create the mental image of Italians in the foreign audience.

The language used by the characters in the original English version is a hot topic in recent articles discussing the film.[20] It seems to me that there is a big linguistic difference between the inhabitants under the water's surface and the locals in Portorosso. The former tend to use standard North-American English, while the people of Portorosso tend to adopt a kind of Italian English that, in line with Kozloff's function of "adherence to the diegesis", is responsible for creating the fictional world of the narrative. This Italian English variety is completely unrealistic and deviates from the norm in pronunciation and the use of Italianisms. What strikes the viewer, however, is

[20] See, for example, Clarke (2021); Hogarty (2021); NPR (2021).

the portrayal of Italians speaking English – a language different from Italian. The opening scene not only sets the spatial and temporal framework for the story – as already described – but also establishes the linguistic variety that the viewer will experience throughout the film. The fishermen Tommaso and Giacomo speak English with a strong Italian accent, which is strange as there seems to be no reason why two old, Italian fishermen in Liguria would do so. The variety used by the people of Portorosso, referred to in this article as Italian English, does not aim to realistically reflect the way people would speak in the Cinque Terre in the mid-1950s. Kozloff′s function of "adherence to the code of realism", which aims to make the dialogues sound realistic, despite being perfectly adaptable to accented voices, as suggested by Ranzato (2021), is rather disregarded in *Luca*.

The Italian English variety certainly aims to anchor the characters in the diegesis, but the function of film dialogues that is most used in *Luca* is "exploitation of the resources of language", where the audiovisual language is anything but realistic, being rather an artificial variety that creates a comic effect. The way inlanders pronounce English is certainly comical for English speakers, and the same goes for Italians when it comes to Italianisms, which are often mispronounced or creatively invented as if the characters were foreigners and not from Liguria. Both the varieties (i.e. the standard North-American English used by the sea monsters and the Italian English used by the people of Portorosso) are rather informal and colloquial as well as anachronistic, since the English slang words used in the film do not fully correspond to the years in which the story is set. Moreover, it is not surprising that the two protagonists of the film belong to the underwater world, where the standard language is spoken; as a matter of fact, Italian English is mainly used to portray minor characters (who are, however, consistently shown on screen), such as Ercole Visconti, Massimo Marcovaldo, and other passengers like a priest, a policewoman, fishermen, and old men and women on the street. They play a more or less secondary role in the film, with the exception of Giulia Marcovaldo, who could be considered a co-protagonist. Despite the use of Italianisms and typical features of Italian English, Giulia′s accent is less strong than the others′, perhaps due to her more central role in the film, or her stay in Genoa for her studies. Had the protagonists spoken this fake English variety full of Italianisms all the time, it might have been more difficult to follow the story and the audience would have struggled to empathise with these characters; the reason for this is what is known as "reader resistance", which is perfectly adaptable to audiovisual texts – caused by "rendered speech that departs to any appreciable degree from standard colloquial speech" (Toolan 1992: 34). Surprisingly, there is no evidence of Italian dialects or non-standard Italian accents in the English film, which is unexpected given the chronological and social setting (i.e. mainly working-class fishermen in the 1950s). It is hard to believe that mainly old people in a small town in Italy in the mid-1950s would speak standard Italian and not non-standard dialects. The Italian lexicon used in *Luca* is not dialectal, but rather belongs to what Sobrero – Miglietta (2011: 99) call "italiano popolare",[21] "quell'insieme di usi frequentemente ricorrenti nel parlare e (quando sia il caso) nello scrivere di persone non istruite e che per lo più nella vita quotidiana usano il dialetto, caratterizzati da numerose devianze rispetto a quanto previsto dall'italiano standard normativo."[22]

The Italian English variety is characterised by the following phonological and prosodic features, partly adapted from Mammen – Sonkin (1936) and used throughout to represent the Portorosso people in *Luca*. In particular:

(a) Vowels

- Because Italian has fewer vowels than English does (7 compared to 20), and certain vowel substitutions occur here, this variety shows a reduction in the number of vowels used, thus [i:] for [i:] and [ɪ], [u:] for [u:] and [ʊ], etc.; moreover, speakers of Italian pronounce some English vowels with greater quantity (length);
- Certain diphthongs show monophthongization of [eɪ] to [e:], and [ou] to [ɔ:];
- Occasional paragoge of the vowel schwa [ə] results in the addition of this vowel to the ends of consonant-final English words, since Italian words are regularly vowel-final;

(b) Consonants

- [r] is pronounced and trilled in all positions, especially inter-vocalic ones;
- [Θ] and [ð] are pronounced as [t] and [d];
- The plosives [p] and [k] can be dentalised and unaspirated;
- In initial, prevocalic position, [h] is dropped, as in Italian;

[21] Popular Italian.

[22] Linguistic uses that are typical of the spoken and (sometimes) written language, common among uneducated people who mainly use dialect in daily life, and characterised by numerous deviations from standard Italian. (author's translation)

(c) Prosody

- Intonation exhibits a pitch range which is wider than it is in English;
- Suprasegmental patterns differ from those of English, and syllable timing, regular in Italian, can replace stress timing, which results in increased stress on syllables receiving secondary or tertiary stress in English.

From a lexical perspective, code-switching is consistent throughout the film. Code-switching "refers to instances when speakers switch between codes (languages, or language varieties) in the course of a conversation. Switches may involve different amounts of speech and different linguistic units – from several consecutive utterances to individual words and morphemes" (Swann et al. 2004: 40). Code-switching in *Luca* occurs mainly inter-sententially, i.e. a switch occurs at the end of a sentence/clause-level unit and marks the unit that follows. However, there are also cases of intra-sentential code-switching – also known as code-mixing – which "involves the embedding or mixing of various linguistic units […] from two distinct grammatical systems or subsystems within the same sentence and the same speech situation" (Tay 1989: 408). Inter- and intra-sentential code-switching thus signal the different identities with which a speaker is endowed and which are reflected in (or rather constructed by) the different codes s/he uses. In *Luca*, however, this does not seem to be the case. Code-switching does not signal that the people of Portorosso can speak both English and Italian, but is a fictional construction to convey the idea that the people of Portorosso are Italian and, in a strange agreement between the producers and the audience, must restrict their Italian to certain situations and use English more extensively in order to be understood by the English-speaking audience. The use of English is only functional for understanding, and the true identity of the people of Portorosso is revealed when they speak Italian – their "real" language. Code-switching does not occur randomly, and many situational variables and grammatical rules influence the frequency and position of code-switching. In *Luca*'s case, for instance, code-switching seems to occur more frequently when the speaker experiences an emotional shift, often but not necessarily for face attacking purposes. The expression "emotional shift" has been adapted from Hodson's "emotional style-shifting", which, in contrast to code-switching, refers to a change between speech styles *within a single language* (my emphasis) caused by a sudden change in the speaker's emotional state. Emotional style-shifting occurs when characters are surprised, upset or disturbed from their normal emotional state; the same is true of code-switching, which, unlike style-shifting, is an inter-linguistic phenomenon that occurs when characters switch from one code to another. Through "emotional" code-switching, speakers show their true nature, because speech styles expressed when people are under emotional pressure seem to be more authentic (Hodson 2014: 174-175).

The Italianisms in *Luca* have four main functions, which are quite well balanced, as can be seen in *Figure 1*.

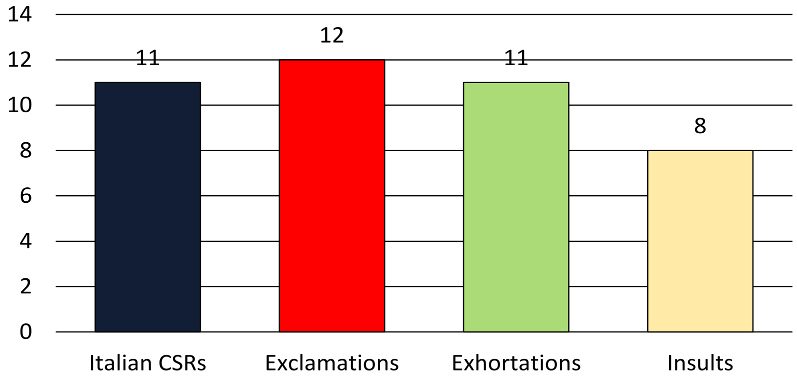


Figure 1. Italianisms in Luca

They occur more frequently in exclamations (29%), which generally express a sudden emotional shift. As can be seen in *Table 1*, most exclamations do not exist in Italian and are constructed on the basis of typical Italian food ("per mille sardine", "per mille cavoli", "santa mozzarella", "santa ricotta", "santo pecorino", "santo gorgonzola"). They replace common Italian expressions that contain religious elements that would not be understood by the English-speaking audience (e.g. "santo cielo", "santi numi", "santa madre") with typical Italian food that is well-known abroad and in most of the cases cannot be translated into

Table 1. Exclamations

Exclamations
"*Per mille sardine*!"
"*Mannaggia*, here we go!" "*Mannaggia*, not a great catch today!"
"*Santa mozzarella*, we did it!"
"*Santo pecorino*, that's the best idea ever!"
"*Porca paletta*, what was that?"
"*Per mille cavoli, Guido*!"
"*Mamma mia*!" "*Oh mamma mia*, please no more ravving!" "*La mia bambina*! *Oh mamma mia*!"
"*Santa ricotta*!"
"*Oh santo gorgonzola*, I need to pack for school!"

English (e.g. mozzarella, pecorino, ricotta, gorgonzola). These exclamations are mainly used to express surprise (e.g. a fisherman shouts "per mille sardine!" after seeing sea monsters; Luca exclaims "santa mozzarella!" after riding a bike for the first time; Giulia exclaims "santa ricotta!" after finding out that Luca and Alberto are sea monsters), which is also expressed with the exclamation "porca paletta!". The interjection "mannaggia",[23] which also occurs in Italian, is used to express bother (as in "mannaggia, here we go!" exclaimed by a policewoman when she hears Ercole's noisy Vespa approaching) and regret ("mannaggia, not a great catch today!", exclaimed by Massimo Marcovaldo). Fear is expressed above all with "mamma mia", when Ercole is afraid of the sea monsters or his sparkling Vespa falls down. "Mamma mia" also expresses exhaustion as in "mamma mia, please, no more raving!".

Italianisms are also used for Italian culture-specific references (CSRs, 26%, see *Table* 2), i.e. "words or composed locutions typical of a geographical environment, of a culture, of the material life or of historical-social peculiarities of a people, nation, country, or tribe and which, thus, carry a national, local or historical colouring and do not have precise equivalents in other languages" (Ranzato 2015: 67). In *Luca*, Italianisms are used to express mainly ethnographic references,[24] more specifically objects of daily life (pescheria,[25] trenette al pesto, pasta, fusilli, trofie, cannelloni, lasagne, espresso, olio d'oliva[26]). There is a case of socio-political CSR (Maggiore[27]), which refers to institutions and functions. Most CSRs refer to typical Italian food and drinks that are well known all over the world. CSRs borrowed from a foreign language are useful for constructing an exotic environment, as they convey an air of foreignness.

Table 2. Italian culture-specific references

Italian culture-specific references
"It smells like behind the *pescheria*"
"*Maggiore*, another sighting, in the harbour this time"
"Dinner's ready. *Trenette al pesto*"
"Every year they change the *pasta*. You have to be ready for everything. Could be *cannelloni*, *penne*, *fusilli*, *trofie*, even *lasagne*"
"*Espresso*!"
"Ciccio, hold still. *Olio d'oliva*"

[23] Damn!

[24] For a classification of CSRs, see Díaz Cintas – Remael (2007: 201).

[25] Fishmonger.

[26] Olive oil.

[27] Major.

Table 3. Exhortations

Exhortations
“*Andiamooooo*!” “Stop crying and tag Guido. *Andiamo*!”
“*Silenzio, Bruno*!”
“*E basta*!” “Hey, *Ercole, basta*!”
“*Mangiamo*!”
“*Forza, Luca*!” “*Forza, Giulietta*!”
“*Buongiorno, andiamo dai*!”
“*A casa*!”
“Out of the way, *via, via*!”

Exhortations and orders, as shown in *Table 3* (26%), imply an emotional and power imbalance, where speaker A imposes his/her decision on speaker B. They are examples of intentional FTAs directed at speaker B’s negative face. Brown and Levinson (1987) claim that FTAs addressed to the speakers’ negative face (i.e. the desire not to be hindered in one’s actions) take the form of an order, a request. This is the case with imperatives such as “andiamo!”,[28] “mangiamo!”,[29] “via, via!”,[30] “a casa!”[31] (in this last example, the verb “andiamo” is omitted), but also of the interjection “basta!”,[32] with which an old woman rebukes a group of noisy boys. The interjection “forza!”[33] is used instead to support the listener – not the opposite. “Silenzio, Bruno!”[34] is an FTA against Bruno’s negative face, an imaginary voice in Luca and Alberto’s heads – a kind of conscience – that clips their wings; for this reason, it should be silenced.

Table 4. Insults

Insults
“What’s wrong with you, *stupido*!”
“*’sto imbecille* thinks he can be a jerk” “*Imbecille*!”
“You can’t swim, you can barely wide a bike. *Siete un disastro*!”
“*Ma sei scemo, Ercole*!?”
“*Disgraziati*!”
“Ah, *idioti*, you let it get away!” “Eat, *idiota, più veloce*!”

Furthermore, Italianisms are used to express insults (*Table 4*, 19%), which, unlike exhortations, are FTAs against people’s positive face, i.e. the desire to be recognised. The insults are mainly voiced by Giulia and Ercole, both very loud characters (Giulia complains

[28] Let’s go!
[29] Let’s eat!
[30] Go away!
[31] Go home!
[32] Enough!
[33] Come on!
[34] Silence, Bruno!

that people think she is "too much"). Giulia's insults are mainly directed at Ercole's positive face ("imbecille",[35] "scemo"[36]), while Ercole's insults are directed at his supporters Ciccio and Guido ("disgraziati",[37] "idioti"[38]).

It is interesting to note that the use of Italian is often associated with impoliteness. This is because most Italian characters in *Luca* are portrayed as extremely dynamic, sociable and passionate people who tend to talk a lot, loudly and expressively. However, the use of impolite Italianisms should be seen as a natural consequence of a change in the emotional status of the passionate Italian characters, who switch to the language "of the heart" when they feel the need to express something heartfelt. This is common among people who speak more than one language, one of which (or more) tends to have affective connotations and is considered "better" for expressing a person's emotional status.

5. Conclusions

In 1940, Walt Disney Productions released *Pinocchio*, an American animated musical fantasy film based on the 1883 Italian children's novel *The Adventures of Pinocchio* by Carlo Collodi. This was the second animated film produced by Disney (after *Snow White and the Seven Dwarfs*, 1937), and the first (and last) film set entirely in an Italian village (in Tuscany) and featuring only Italian characters. Most of the characters' names are Italian (e.g. Geppetto, Pinocchio, Figaro, Cleo, Stromboli). However, only one of the characters, Stromboli, speaks English with a strong Italian accent. He is a cruel puppet-maker who forces Pinocchio to perform in his theatre to earn money and uses him as firewood when he grows old. He exemplifies Disney's dishonest villain. Despite his Italian accent, there is no sign of Italianisms in the language used to portray Stromboli. When emotional shifts occur (especially when he gets angry), he speaks slurred words with a typical Italian prosody and sounds. The construction of Italianness in *Pinocchio* is thus minimal compared to that in *Luca*. This could be due to the different trends in the representation of foreign characters in the two eras in which these products were released – i.e. the 1940s and the 2020s, respectively.

[35] Imbecile.

[36] Fool.

[37] Rotten.

[38] Idiot.

Eighty-one years after *Pinocchio*, Disney and Pixar are back with a new animated film set entirely in Italy. Unlike *Pinocchio*, *Luca* is a tribute to Italy and its culture. The visual representation of a small coastal town in the Cinque Terre in the mid-1950s is meticulous, and the language adopted is worth studying. The producers put extensive effort into creating an artificial language that would convey the idea of exoticism in both time and space. As mentioned earlier, it is an English-based variety that differs from standard North-American English in both its pronunciation and lexicon. The Italian accent is used to characterise only the inhabitants of Portorosso and to distinguish them from the sea monsters living under the water's surface. The accent is stronger in Ercole Visconti, who embodies the loud and boastful Italian bully, who is rich and ostentatious, and weaker in Giulia Marcovaldo, who lives in Genoa, where she goes to school, and only returns to Portorosso in the summer, thus losing some of the "rusticity" of Portorosso locals. Unlike in *Pinocchio*, where Stromboli stammers, confusing Italian sounds that are incomprehensible to both English-speaking audiences and Italians, the people of Portorosso wrap up their sentences either with real Italianisms or with creative expressions that do not exist in Italian, but are perfectly understandable to both English-speakers and Italian-speakers. These expressions make consistent use of typical Italian food, well-known all over the world. As described in previous sections, characters switch to Italian mainly in response to emotional outbursts, as evidenced by the high frequency of Italianisms in exclamations, exhortations and insults. In addition, Italianisms are also used for CSRs, especially to refer to food and drink. Italian expressions are standard but belong to a low register (*Italiano popolare*), characterised by colourful expressions, vernacular imprecations ("mannaggia"), and apheresis, as in "'sto imbecille", where the adjective "questo" (this) is reduced to "'sto". Interestingly, no Italian dialects appear, which could be explained by commercial reasons behind the American production of the film. Paradoxically, the effort made by *Luca*'s producers to distinguish the language of the sea monsters from that of the people of Portorosso is unfortunately lost in the Italian dubbing, where dialects could be used for characterisation. Nevertheless, all the characters speak standard Italian indistinctly and the funny moments created by the use of Italianisms in the original are eliminated. While it is true that the portrayal of people of Portorosso is more authentic in the Italian dubbed version, since they are Italians who actually speak Italian, perhaps the Italian dubbing could have used accents and dialects from Liguria to distinguish the inland characters from those who live underwater, as is done *mutatis mutandis* in the original version.

REFERENCES

Sources

Luca
2021 Directed by E. Casarosa, Pixar Animation Studios and Walt Disney Pictures.

Pinocchio
1940 Directed by N. Ferguson – T. Hee – W. Jackson. Walt Disney Productions.

Snow White and the Seven Dwarfs
1937 Directed by W. Cottrell – D. Hand – W. Jackson. Walt Disney Productions.

Soul
2020 Directed by P. Docter – K. Powers, Pixar Animation Studios and Walt Disney Pictures.

Special studies

Brown, P. – S.C. Levinson
1987 *Politeness. Some Universals in Language Usage*. Cambridge: CUP.

Collodi, C.
1883 *Le Avventure di Pinocchio*. Giunti Junior.

Clarke, D.
2021 "Luca: Pixar's new film is lively and kind of funny. But it's no Finding Nemo", https://www.irishtimes.com/culture/film/luca-pixar-s-new-film-is-lively-and-kind-of-funny-but-it-s-no-finding-nemo-1.4593961, accessed July 2021

Díaz Cintas, J. – A. Remael
2007 *Audiovisual Translation: Subtitling.* Manchester: St Jerome.

Ferguson, S.L.
1998 "Drawing fictional lines: Dialect and narrative in the Victorian novel", *Style* 2, 1-17.

Gross, L.
1991 "Out of the mainstream", *Journal of Homosexuality* 21 (1-2), 19-46.

Hall, S.
1997 *Representation. Cultural Representations and Signifying Practices.* London: Sage.

Hodson, J.
2014 *Dialect in Film and Literature*. Basingstoke: Palgrave Macmillan.

Hogarty, J.
2021 "REVIEW: "Luca"; A movie with a lot of potential that just never gets off the ground", https://wdwnt.com/2021/06/review-luca-a-movie-with-a-lot-of-potential-that-just-never-gets-out-off-the-ground/, accessed July 2021

Jefferson, C.

2021 "Exploring friendship, acceptance, and overcoming fear in Pixar's Luca", https://news.disney.com/luca-first-look, accessed July 2021

Kozloff, S.

2000 *Overhearing Film Dialogue*. Berkeley, CA: University of California Press.

Lippi-Green, R.

2012 *English with an Accent: Language, Ideology, and Discrimination in the United States*. London: Routledge.

Mammen, E.W. – R. Sonkin

1936 "A study of Italian accent", *Quarterly Journal of Speech* 22 (1), 1-12.

McConnell-Ginet, S.

2011 *Gender, Sexuality, and Meaning: Linguistic Practice and Politics*. Oxford: OUP.

Montini, D. – I. Ranzato

2021 "Introduction: The dialects of British English in fictional texts: Style, translation and ideology". In: D. Montini – I. Ranzato (eds.) *The Dialects of British English in Fictional Texts*. New York; London: Routledge, 1-8.

Motschenbacher, H.

2011 "Taking Queer Linguistics further: Sociolinguistics and critical heteronormative research", *International Journal of the Sociology of Language* 212, 149-179.

Nemiroff, P.

2021 "What do Pixar sea monsters look like? 'Luca' – director Enrico Casarosa explains", https://collider.com/luca-sea-monsters-explained-enrico-casarosa/, accessed July 2021

NPR

2021 "In 'Luca', the tears just might sneak up on you", https://www.npr.org/transcripts/1006824991?t=1638780241448, accessed July 2021

Pavesi, M. – M. Formentelli – E. Ghia

2015 "The languages of dubbing and thereabout: An introduction". In: M. Pavesi – M. Formentelli – E. Ghia (eds.) *The Language of Dubbing: Mainstream Audiovisual Translation in Italy*. Bern: Peter Lang, 7-26.

Ranzato, I.

2015 *Translating Culture Specific References on Television: The Case of Dubbing*. London; New York: Routledge.

2021 "The accented voice in audiovisual Shakespeare". In: D. Montini – I. Ranzato (eds.) *The Dialects of British English in Fictional Texts*. New York; London: Routledge, 147-164.

Ranzato, I. – Zanotti, S.

2018 "Introduction: If you can't see it, you can't be it: Linguistic and cultural representation in audiovisual translation". In: I. Ranzato –

S. Zanotti (eds.) *Linguistic and Cultural Representation in Audiovisual Translation*. New York; London: Routledge, 1-8.

Sobrero, A.A. – A. Miglietta
2011 *Introduzione alla linguistica italiana.* Bari: Laterza.

Style
2021 “Is Luca Pixar's first gay movie? How the Disney+ film's ‘deeper story’ and animation design came together, with a little help from Renaissance maps and sea iguanas”, https://www.scmp.com/magazines/style/leisure/article/3138255/luca-pixars-first-gay-movie-how-disney-films-deeper-story?module=perpetual_scroll&pgtype=article&campaign=3138255, accessed July 2021

Swann, J. et al. (eds.)
2004 *A Dictionary of Sociolinguistics.* Edinburgh: Edinburgh University Press.

Tay, M.W.J.
1989 “Code switching and code mixing as a communicative strategy in multilingual discourse”, *World Englishes* 8 (3), 407-417.

Toolan, M.
1992 “The significations of representing dialect in writing”, *Language and Literature* 1 (1), 29-46.

Trudgill, P.
1994 *Dialects*. London; New York: Routledge.

Address: Davide Passa, Sapienza Università di Roma, Dipartimento di Studi Europei, Americani e Interculturali, Piazzale Aldo Moro 5, 00185 Roma, Italy.
ORCID code: orcid.org/0000-0003-3327-2101

Token: A Journal of English Linguistics 13, 2021

"Simplicity is the ultimate sophistication" or half a century of IT consumer identity formation: A pragmatics approach

Nataliia Kravchenko*, Olga Valigura*, Vira Meleshchenko**, and Liudmyla Chernii**

* *Kyiv National Linguistic University*
** *Ternopil Volodymyr Hnatiuk National Pedagogical University*

ABSTRACT

The article examines the genesis and modification of IT consumer's identity (ITCI) in terms of certain pragmatic properties of Apple's slogans. Drawing on Barthes's concept of mythologization, underpinned by theories of personal archetypes and Maslow's hierarchy of needs, the study identified ITCI-descriptors – stylistically and pragmatically connotated meanings, associated with the advertised product or customer characteristics, related to ITCI formation.

Initial ITCI construction relies on cognitive needs and the Explore archetype, based on customer-associated descriptor "creativeness", marked by disregard for cooperative maxims, by oxymorons, allusions, puns and aposiopesis, iconically reproducing non-standard thinking. Subsequent stages involve the Seeker archetype hybridization with the Trickster archetype, related to ludic stylistics, paradoxes, non sequitur, and occasionalisms-compounding. Currently creativity-based identity gives way to universalization-based ITCI marked by positive politeness, indirect commissives, pronouns of inclusiveness, indefiniteness, and metonymic identification of the product with its owner. Product-associated descriptors are at the core of the ITCI field of needs. Peripheral is the need for respect, even more peripheral is the need for in-group affiliation and cognitive needs.

Keywords: IT consumer's identity, advertising slogan, historical development, pragmatics, stylistics.

1. Introduction

The problem of formation and transformation of consumer identities, negotiated within social, technological, globalization processes remains

DOI: 10.25951/4824

of the upmost interest for interdisciplinary and linguistic studies since it contributes to the complex of theoretical issues of discursive, narrative, semiotic, symbolic, interactive, multimodal mechanisms of the identities construction and manifestation and in the applied sense – for tracking and predicting the dynamics of consumerism and its impact on the motivational structures of consumer identities, including in the transnational perspective. In contemporary studies such identity is specified primarily as a discursive construct, which changes with the rearticulation of discourse meanings (Hammack 2008: 2) as well as continuous (Cherrier – Murray 2007; Elliott 2004), interminable (Gabriel – Lang 2006), narrative (Ahuvia 2005; Marion – Nairn 2011; Mikkonen et al. 2011) and symbolically projective (Mikkonen et al. 2011) phenomena. That is, different approaches to the study of identity emphasize its dynamic, changeable nature, dependence on global processes of society, discourses and culture.

The problem of the genesis and historical development of IT consumer identity (ITCI) in the tangible time frame of the last fifty years is associated with a paradigmatic rethinking of linguistic approaches to the study of identity.

Viewed from a cognitive-discursive approach in its semiotic framework, the construction of consumer identity is not so much influenced by linguistic variations and sociolinguistic variables such as social class, gender, and space within speech communities, as by conceptual rearticulations within the discourses, which create their target identity. Discourses promoting transnational technologies focus on creating transnational consumer identity. One of them is the Apple advertising discourse, which over the past half century has been constructing its target transnational consumer with the involvement of such universal human resources as unexpected, often provocative stylistic and pragmatic devices, universal narratives and intertextuality, symbolic archetypal codes, music, and other multimodal resources.

The evolution of such an identity, specified from a linguistic and pragmatic point of view is the main focus of this study.

The novelty of the research lies in the analysis of the slogans of Apple's multimodal advertising discourse from 1976 to the present from the point of view of their pragmatic and stylistic characteristics that reveal strategies for the formation of a "project" consumer identity with an emphasis on conflict, strengthening, weakening, and development of ITCI structural elements at different stages of its actualization. An additional perspective of the research is the interdisciplinary analysis of the identified components from the point of view of the corresponding personality archetypes in projection on the dynamics of ITCI motivations based on Maslow's hierarchy of needs.

2. Literature review

The article integrates a pragmatic and stylistic analysis of Apple's slogans with Barthes's (1973) concept of the second-level signification and mythologization, in terms of the formation of consumer values by Apple's discourse. The values that make up the second level of the signification-mythologization of the Apple discourse are analyzed in two dimensions: from the point of view of their encoding by verbal and pragmatic markers and in the discursive and semiotic aspect – in connection with the construction of consumer identity by appealing to its value motives.

With that in mind, the theoretical-methodological premises of the paper involve (a) the studies which specify the processes of identities construction and manifestation within different research paradigms; (b) the study of discourse-constructed and identity-forming mythologemes associated with IT products, which are "alienated" from these products' functional purpose, being superseded by the target identity values. At this stage of analysis, the paper partially employs the concept of the personal archetypes as patterns of behavior and motivation, consistent with the projected ITCI values and needs, as well as Maslow's hierarchy of needs.

Let us briefly turn to the analysis of each of these research approaches.

A conceptual framework for identity construction from a pragmatics perspective involves two main scholarly strands, which can roughly be termed as "interactive" and "discourse-constructionistic" (semiotic). Within the first interactionist paradigm the identity construction is closely associated with the notion of performance, when identity as membership in various categories is constructed, assumed, attributed or resisted in the process of discursive interaction. In other words, 'constructing identities' is viewed as a kind of social and 'discursive work' (Kravchenko – Pasternak 2018; Zimmerman – Wieder 1970).

The "discourse-constructionistic" approach is more consistent with the specifics of the paper due to the impossibility of tracking interactive feedback from the advertising consumer. This approach views identity as a discursive construction (Bamberg et al. 2011), taking into account the interrelated processes of constructing identities through discourses and the formation of discourses by identity actors in a wide historical and institutional context (Fairclough 1992).

The socio-semiotic perspective (Dunn – Neumann 2016; Hodge – Kress 1988; Kress 2010; van Leeuwen 2005) makes it possible to explain the influence of advertising discourse in its "world-modeling" properties on the creation of consumer identity by constructing, maintaining and transforming its underlying values. In this connection, the social-semiotic approach is consistent with the concept of the second level of signification and mythologization, introduced by R. Barthes (1973) – considering the relationship "product (denotative meaning – signifier) – value meaning attributed to it (connotative signified) – the target identity component associated with connotative signified". According to Barthes, the meaning (first signified) generated by the linguistic code, which designates the goods or service, becomes a form (a signifier) for the new signified – a concept that "alienates" the "natural" function of the goods and replaces the initial denotative meaning with its associating values.

In the third (mythological or ideological) order of signification, a set of constructed signs-connotatums or mythologemes forms a discursive mythology, creating a "worldview, one of the "possible worlds", positioning the modeled reality as objective and non-alternative" (Kravchenko et al. 2020a: 315-316). Such a possible world becomes a picture of the world that determines one or another type of consumer identity.

Based on Erikson's definition, that identity "connotes both a persistent sameness within oneself (selfsameness) and a persistent sharing of some kind of essential character with others" (Erikson 1980: 109) and bearing in mind that archetype motivations can interact with the main components of ITCI, the article uses the concept of personal archetypes as psychologically motivated mental models, such as schemes and prototypes of oneself or others, acting on an automatic or unconscious level (Lindenfeld 2009).

In particular, the research employs the taxonomies of the basic archetypes, borrowed from archetype psychology (Jung 1969, 1971; Pearson 2015) and "neo-archetypal theory" (Faber – Mayer 2009), which are widely used in branding and advertising marketing. Since ITCI motives are ultimately built on main human needs, the paper partly addresses Maslow's hierarchy of needs (1943, 1970a, 1970b).

3. Database and methods

The approach taken in this study introduces and uses the operational unit "identity descriptor" or, in the context of this article, the ITCI descriptor, which is understood as representing a stylistically and pragmatically connotated meaning, referring either to the advertised product property or to the target customer characteristic that motivates the choice of product premised on a certain structural component of its consumer identity.

The material includes Apple's slogans, sampled from the company's advertising discourse from 1976 to the present. For the analysis, 100 slogans that together provide a holistic idea of the Apple strategies for constructing its target consumer identity were selected – taking into account the ITCI structural components and the dynamics of their development and transformation over almost half a century. The main criterion for the selection of material was the criterion of diversification of consumer values and mythologemes encoded with slogans, combined with the criterion of representation of various historical periods of the company's development and corresponding advertising projects. At the same time, slogans, which in different versions signify the same consumer values, were excluded from the sample. Considering that many of Apple's slogans reflect similar values, such slogans were sorted out according to a criterion for combining similar values as hyponyms of one hyperonymic value. For example, slogans *1,000 songs in your pocket*, *Tune your run*, *Clip and go*, etc. actualize the same value – convenience / ease of use, highlighting different aspects of this meaning at the level of contextual connotations.

The integrative method of analysis encompasses (a) methods of stylistic analysis (Simpson 2014) to identify stylistically highlighted connotative meanings associated with the descriptor of either the mythological value of goods / services, or the basic characteristic of the target customer, underlying a specific structural component of ITCI; (b) pragmatic analysis based on Grice's theory of cooperativeness (1975, 1989) and conversational implicature (Bach 2010, 2012; Potts 2015) as well as speech act theory (Austin 1970; Searle 1969) and explanatory tools of the Politeness theory (Brown – Levinson 1987; Leech 2014) with reference to Relevance Theory (Wilson – Sperber 2004; Carston 2004), having in mind some connotative correlations between the pragmatically indexed directness-indirectness, distance or proximity, coherence-incoherence, relevance-irrelevance, etc. and verbally encoded motivational descriptors; (c) intertextual analysis aimed at inferencing the allusion, encoding by slogans, and interpreting the slogan implicature (if available) within the context of the Apple advertising discourse of a particular period; and it incorporates (a) elements of social-semiotic analysis in terms of Barthes's signification levels – to identify mythological properties of the goods associated with a particular component of the project identity; and (b) elements of archetypal analysis to specify the identity descriptors correlation with certain personal archetypes (Faber – Mayer 2009; Jung 1971; Pearson 2015; Shadraconis 2013).

In the present investigation, the interaction between pragmatics and discourse analysis is also of some interest, which is studied using the concept of optimal relevance, taking into account various types of contextual effects. In this respect, the relevant context is the discourse of a particular advertising company, which can enhance, maintain, eliminate, etc., the meaning denoted or connotated by the slogan. Here, the attraction of the optimally relevant context correlates with the concept of intertextual analysis "bridging the gap between texts and contexts" put forward in the framework of critical discourse analysis (Fairclough 1995: 188). Thus, intertextuality is examined in this discussion both in a narrow sense (as a stylistic device) and in a broad sense – as the interaction of the value sense of the slogan with the integral context of an advertising campaign of a certain period. On the other hand, in view of the innovative figurativeness of the Apple slogans, the notions of the optimally relevant context and its pertinent intertextuality are reinterpreted in this paper from the viewpoint of the Optimal Innovation Hypothesis (Shuval – Giora 2005) – with due attention to the scale of stimuli associated with different stages of the Apple discourse and its target identity construction.

The data analysis involves five consecutive stages.

The first stage consists in the identification of the samples of the research material, representing different consumer values of the Apple discourse and being marked by particular stylistic and pragmatic properties.

The second stage involves the pragmatic analysis of the slogans of different temporal spans, revealing the markers of (a) cooperative maxims adherence or disregard which resulted in discursive implicatures; (b) positive or negative politeness strategies; (c) direct / indirect speech acts, and their illocutionary force – to specify the pragmatic properties impact on connoting the meaning of distance, proximity, fidelity, ambiguity, etc. associated with pragmatic facet of the ITCI descriptor. Where the maxims' flout the implicature-associated inference, the hypotheses are verified within the context of the whole discourse of the corresponding advertising campaign in terms of optimal relevance and cognitive effects of strengthening, contradiction, or eliminating an assumption, triggered by insufficiency, untruthfulness, incoherence, intransparency of information expressed by the slogan.

The third stage consists in revealing the stylistic properties of slogans associated with the connotated quality or value that are important for accentuation / construction of a certain type (structural component) of ITCI. Investigating the interface of stylistic and pragmatic means with the characteristics of consumer identities, this study partially draws on research on the iconic or indexical properties of verbal and pragmatic resources (Bordron 2011; Holzscheiter 2014; Kravchenko – Zhykharieva 2020). Herein, we take the view that the isomorphism of a certain structural component of consumer identity may function as a means of its actualization.

In particular, different expressive and pragmatic means are aimed at attracting different psycho-emotional types of people. Non-standard puns, oxymorons or sophisticated allusions are targeted at ITCI, interested in the product-associated creativity. The pragmatics of such slogans presupposes the cooperative maxims flouting and inference of discursive implicature, requiring additional cognitive efforts. On the other hand, laconic, informative messages based on distance-reducing directives and positive politeness strategies are rather aimed at the needs of belongingness to a certain in-group identity consisting of clients and the Apple team.

The fourth stage of research focuses on the interpretation of the identified ITCI descriptors in terms of their associated personal archetypes as patterns of consumer behavior, as well as on the framework of Maslow's hierarchy of motivations.

The fifth stage consists in the generalization and interpretation of the results obtained at the previous stages of the analysis from the point of view of the distribution of stylistic and pragmatic stimuli on the scale of optimal innovation and pleasure - taking into account different parameters of "familiarity".

The sixth stage of the analysis consists in comparing the ITCI descriptors typical for different stages of the construction of consumer identities – in order to identify patterns and trends in articulation, rearticulation, exclusion, etc. structural components of ITCI, formed by the Apple discourse.

3.1 First stage of IT consumer identity formation: Affiliation, creativity, or security?

The first stage of IT consumer identity construction is based on two key customer-associated descriptors, i.e. "creativity" and "in-group affiliation", as well as on the secondary, product-associating descriptors, whose actualization is based on connotations, and is foregrounded by stylistic and pragmatic devices as displayed by the Table 1. A connotated meaning, referring either to the characteristics of a product or to a characteristic of the target customer, is aimed at motivating the choice of a product by appealing to a certain structural component of its consumer identity.

So, descriptors have more to do with the motivational needs of the targeted INCI that unites Apple customers than with the characteristics of the advertised product itself. In other words, the signified meanings associated with the slogans are as follows: Apple IT products symbolize the creativity of their owners; Apple IT products symbolize a special group identity. As the research material has shown, the Apple discourse of the first period of the ITCI formation links these two meanings: creativity is a sign of the group identity of those who own the Apple product.

To create this level of symbolization (mythologization), the employed stylistic and pragmatic devices iconically reproduce non-standard thinking at the level of meaning, form and connotation, which is confirmed by the use of oxymorons, allusions, and discursive implicatures, based on the flouting of the cooperative maxims, as shown in Table 1.

In our opinion, the descriptor "creativity" more accurately conveys the Apple message at the first and, partially, the second stages of constructing the target consumer identity than the descriptors "cleverness" and "intelligence". Various definitions (WB, CELD) of the lexeme "creativity" highlight in its meaning the key seme "novelty", in contrast to the words 'cleverness" and "intelligence", the meanings of which do not contain this seme at the level of denotative or connotative components.

Table 1. Descriptors of the first stage of the ITCI formation: Pragmatic and stylistic facets (1976-1996)

Identity descriptors		
Customer-associated descriptors		
Creativeness		
Slogan	Pragmatics	Stylistic devices
Byte into an Apple. (1970s)	Distance-reducing directive as a marker of positive politeness proximity; Disregard for maxim of manner: discursive implicature: (denotative level) the apple company is attributed to as the computer technology symbol; Connotative (Apple – symbol of creativity).	a pun on the word "bite" (bite) and "byte"

Simplicity is the Ultimate Sophistication. (Late 1970s/1980s)	Oxymoron implicature: 1) technological sophistication is hidden behind apparent ease of use; 2) beauty and elegancy form (aesthetic needs satisfaction)	Oxymoron, intensified by attributive "ultimate"; Intertextual allusion on the Leonard da Vinci's words "All ingenious is simple".
Why 1984 won't be like… '1984'	Disregard for maxim of quantity (tautology) and Manner of information, triggering the implicature: Apple is preparing something special this year	Aposiopesis; allusion to the J. Orwell's novel, which serves as an ostensive stimulus for inference about the upcoming changes, which is reinforced and developed in the context of Apple's discourse, where IBM embodies totalitarianism and tyranny in the computer industry associated with the "big brother" from George Orwell. And Apple, respectively, personifies a new and revolutionary IT Maker
need for "in-group" affiliation		
Soon there will be 2 kinds of people. Those who use computers, and those who use Apples. (Early 1980s)	Disregard for maxim of quantity of information (lack of information explaining why it will be so); Indirect comissive containing an illocutionary promise to construct the "in-group"; Indirect commissive as a positive politeness marker.	Implicit antithesis, parallel anaphoric structures of opposing the in-group (Apple users) and out-group (users of other products). Parcellation to emphasize and specify the subsequent information.
The Computer for the rest of us (1984)	Disregard for maxim of quantity of information. Positive politeness strategy of asserting common ground and assuming reciprocity.	Inclusive "we" as an in-group marker; ellipsis.

Of the 235 million people in America, only a fraction know how to use a computer. Macintosh is for the rest of us.	With the above slogan, most Americans are involved in the circle of "their" Disregard for maxim of quantity of information, marked by the second phrase, triggers the discursive implicature "Unlike other computers, Macintosh is easy to use". In combination with relevant context (preceding phrase) this implicature results in contextual implications: most Americans can use it < the rest of Americans is involved in the designed group of the Apple user 4 Positive politeness strategy of asserting common ground	1) semantic technique of evidentiality based on bringing statistical data; 2) marked theme (in Halliday's terms) as non-coincidence of the beginning of a phrase with a phrasal subject to emphasize a particular information; 3) inclusive pronoun in inclusive construction "for the rest of us"
Individuality, internalization of goods as a personal value		
The Power to Be Your Best (1990)	Positive politeness strategy of asserting the concern for the addressee's wants and needs; Indirect comissive speech act	1) Nominative sentence; 2) capitalization to highlight each word of the slogan as an important semantic component; 3) informal possessive *your* as a form of simulated personal address
What's on your PowerBook is YOU, 1992	Disregard for the maxims of quantity and manner of information resulted in implicature: "our product is a part of you"; Positive politeness strategy of presupposing the addressee's knowledge	Capitalization, Inversion, YOU as simulated personal appeal; Visual video support as a set of personal narratives about irreplaceability of PowerBook

product-associated descriptors		
multifunctionality, simplicity, availability		
It does more, It costs less. It's that simple" 1993	Adherence to the maxims of information in explicating the goods characteristics; Indirect commissive as a positive politeness marker.	Simple laconic sentences with anaphoric beginning, expanded by one member – an attribute denoting the advertised quality, reinforced by the antithesis.

With that in mind the choice of the descriptor "creativity" can be supported by the set of arguments. First, the primary version of the logo - with the image of Isaac Newton, and a quote from William Wordsworth, "Newton … a mind forever voyaging through strange seas of thought", written on the frame of the logo, foregrounds the idea of creativity more than any other ideas since Newton and "his" apple is associated more with the idea of a revolutionary discovery in science than with science itself. Second, the rapid evolution of the logo into the polysemantic and ambivalent image of a bitten apple associates not so much with knowledge (of good and evil) as with temptation (in our case, with temptation of the customer for a new revolutionary product). Even if we use the first interpretation (the fruit of sacred knowledge), then such a metaphorical association of revolutionary IT technologies with the forbidden fruit of knowledge is a manifestation of the advertiser's creativity, respectively aimed at a client with a corresponding motivational need. Third, the "rainbow version", of Apple, which iconically refers to the world's first computer with a color display, also emotionally connotates the meaning of creativity. Fourth, creativity is manifested by linguistic and pragmatic devices, as shown by the Table 1, including a pun (*bite/byte*), associating *a bite* with *a byte* - metonymical representation of a tech company, the use of oxymoron, aposiopesis, tautology, resulting in flouting of the cooperative maxims and triggering of discursive implicatures. Finally, a further development of the idea inherent in the descriptor "creativity" is one of the key slogans of the second stage of the Apple consumer identity formation - *Think different*.

At the same time, a strategy of in-group affiliation is implemented not only through the tactics of specialness (associated with creativity), but also through the opposite tactics of universalization – with the aim of expanding the "inner group" at the expense of ordinary people, in addition to the creative and non-standard-minded customers. The tactics of universalization are carried out by ascribing to the goods the connotative meaning of "personal value" associated with its individualization-interiorization in the value system of the identity of the target consumer: a product for any taste, for any purpose, for every person – from a housewife to a professional to a celebrity.

The tactics of universalization via individualization rely, respectively, on the verbal markers of both inclusiveness (the pronoun *we* in different cases and in the possessive form: *Macintosh is for the rest of us*) and individualization (the informal *you* and its possessive form as a manner of simulated personal address: *The Power to Be Your Best*), realizing the fourth (Use in-group markers), fourteenth (assume or assert reciprocity) and ninth (assert, presuppose addressee's knowledge and concern for addressee/s wants) positive-politeness strategies. The main pragmatic result at this stage of ITCI construction is the product signification as a metonymic manifestation of the personality of its owner: *What's on your PowerBook is YOU*. For this purpose, the advertising company additionally uses multimodal narratives of "real" people, celebrities, a certain fictional family, thanks to which the product is attributed the property of an irreplaceable "part" of the personalities of those for whom it is intended.

To a much lesser extent, at this stage, strategies for constructing consumer identity are presented by ITCI descriptors, highlighting the properties of the product itself. Such strategies are pragmatically marked by adherence to the maxims of information in explicating the goods' characteristics, and stylistically by partial parallelism with anaphoric subject and one additional member of the sentence to emphasize the advertised characteristics of the product: *It does more, It costs less. It's that simple.*

The comparison of the early stage of ITCI formation with Maslow's hierarchy of needs shows the priority of the need for the exploration and the search for new meanings over other motivations. Presented in terms of the field structure, identity motivations are distributed as follows: the core values of the targeted identity are the exploration and search for new meanings; the near peripheral zone close to the core is the belongingness need of being part of the consumer in-group; and the far peripheral zone is the safety need, associated with the goods' descriptors, suggesting comfort, security and stability. In other words, Apple's advertising campaigns of the first period primarily involve one of the highest (the fifth of eight) levels of the human needs hierarchy (See Fig. 1).

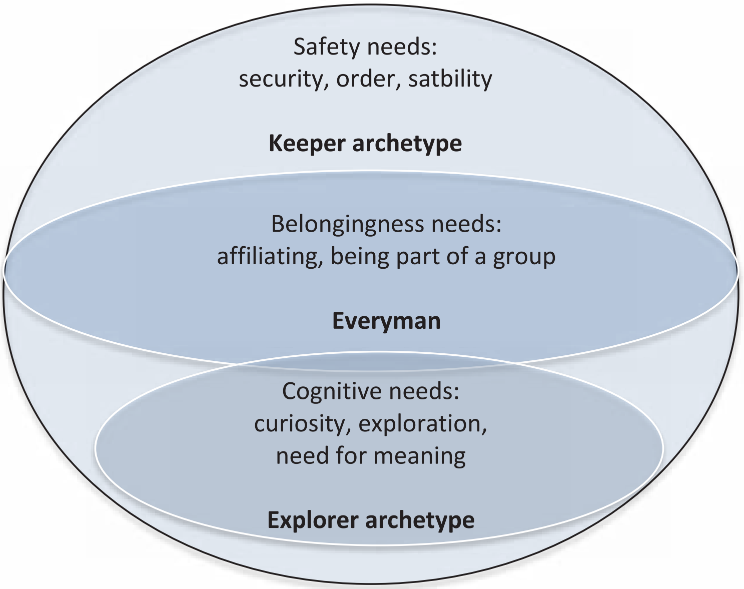


Figure 1. The first stage of ITCI identity formation: Components-descriptors in the field structure of needs

Specification of the identified markers of descriptors within the framework of the Optimal Innovation Hypothesis showed that the creativeness-based figurative means occupy the first places at the scale of the customers "pleasure"-associated attractiveness assured by the familiar intertextuality references to the G. Orwell's novel, the Leonard da Vinci's words, as well as due to the effect of visual iconicity of the product and its pertaining slogan (elegant simplicity of design suggesting sophisticated built-in features form) and allusion to the well-known logotype. The second scale relates to the verbal and pragmatic means, which foreground the in-group identity of the Apple customers. The salience of these mini-narrative-like slogans is marked by disregard for cooperative maxims, which triggers the search of implicature. Its inference is facilitated by the optimal relevant context of positive politeness - inclusive and establishing common ground. The third scale refers to the means related to the product-associated descriptors. They are familiar and easy to decode but lack figurativeness and novelty.

When compared with the classification of archetypes, which are increasingly used in branding practice, we can conclude that the model of customer behavior associated with the search for the new and creative (the fifth level of Maslow's pyramid) is most closely related to the Explorer archetype (Search for yourself, development, discovery of secrets, individuality). The model associated with the universal need to belong to a group is connected with the image of the Everyman (representing belonging, connecting). Finally, the search for comfort, practicality, and convenience (the second level of needs) is linked to the Keeper archetype.

Subsequent stages of ITCI construction either develop and reinforce or eliminate the identity components formed in the first stage.

If we draw a comparison derived from relevance theory to characterize these processes, then we can say that the entire discourse of Apple in its synchronous-diachronic, prospective-retrospective dimensions, in the totality of all verbal and non-verbal signs takes the form of relevant contexts for weakening or strengthening descriptors associated with ITIC components foregrounded by Apple's discourse of the first two decades.

3.2 Second and third stages of the IT consumer identity construction: From complexity to simplification

The second conventionally chronological stage of the Apple ITCT construction, despite the declared message "Think different", which applied to the entire advertising campaign of 1997-2002, focuses equally on the customer creativity ("otherness", "specialness") and their "common sense" ideology associated with the products convenience, comfort and affordability. Accordingly, the first vector of identity formation addressed the qualities of the projective consumer and the second – the characteristics of the IT product, see Table 2.

Table 2. ITCI-descriptors modification: Pragmatic and stylistic devices (1997-2002)

Identity descriptors		
Otherness, specialness, creativeness		
Slogan	Pragmatics	Stylistic devices
Think different, 1997	On-record distance-reducing directive, aimed at "in-group" constructing; Disregard for cooperative maxim of quantity and manner – triggered implicature (exclusivity, uniting the owners of the product)	Intertextuality device – allusion to IBM slogan "Think"

	Allusion-based discursive implicature: we are different from other IT companies and our client should think differently	
iThink, therefore iMac, 1998	Disregard for cooperative maxims of quantity and manner – triggerring first level (focused on product cognitive capacity) and second level (focused on its owner) implicatures	1) Personification; 2) Allusion to Rene Descartes' "I think, therefore, I exist"; graphostylistic device
Hello. Again. 1998	Disregard for the cooperative maxims of Manner, Relevance and Quantity of information, triggering a search for discursive implicature	Personification Parcellation, ellipsis.
product-associated descriptors		
convenience, comfort, economy, power		
Blows minds, not budgets 1998	Disregard for quality maxim, metaphorical implicature: high technological characteristics and availability	Metaphor
The iMac to Go, 1999	Explicature: compact, easy to carry	Ellipsis
Two Brains are better than one, 2000	Explicature: operating system comparable to human cognitive capabilities	Personification, metaphor
1,000 songs in your pocket, 2001	Explicature: significantly smaller and more poweful than competing MP3 players at the time.	metonymic hyperbole
Introducing the new iPod family, 2002	Explicature: novelty and continuity.	personification

The components "creativity" and "otherness" are actualized with practically the same arsenal of pragmatic and verbal means as at the first stage of ITCI construction: (a) the discursive implicatures, which are based on a disregard for cooperative maxims and are inferred in the relevance context of the Apple corresponding campaign discourse: *Hello. Again*. Here at least three cooperative maxims are flouted: the maxim of Manner (where it is unclear who exactly is greeting – the company or the product.), the maxim of Relevance since the semantic coherence and formal cohesion between the parts are violated; and the Maxim of quantity: the slogan lacks information about what exactly happens "again". As a result, a search for implicature is triggered to fill the semantic gap in the relevant context of the advertising campaign ("think differently"), with the inference of the "nearest" meaning: "something fundamentally different has appeared on the Apple market again"; (b) the distance-reducing directive as a marker of positive politeness strategy (or in other words, involvement strategy (Scollon – Scollon 1983: 170), which, together with an allusion hinting at the out-group, serves as a means of in-group construction: *Think different*; (c) the allusions and parcellation still prevail among stylistic devices, while oxymoron and other types of antithesis are not represented.

At the same time, we did not identify the indirect commissives as the markers of the tenth positive-politeness strategy "offer, promise". With respect to the lexical facet all descriptor-based words incorporate denotative meaning "think", which is quite consistent with the main concept of the advertising campaign "think different".

A new aspect in the disclosure of ITCI descriptors "otherness" (creativity) is the attribution of cognitive abilities to the product itself through the stylistic device of personification: *iThink, therefore iMac.* Personification, combined with allusion, becomes a means of characterizing both the product and its owner, precisely due to the relevance context of the campaign ("think different") and to the mythological logic of the extension of the properties of the advertised thing to its owner.

The second ITCI component, that is the consumer "common sense" is focused on IT products' characteristics. From a syntactic point of view, the slogans, which express IT product-associated descriptors are mainly elliptic, concise, nominative, one-part (predicate) sentences highlighting the advertised operations that the product is capable of performing or ease of its use (*The iMac to Go*). Sometimes they incorporate metaphoric components (*Blows minds, not budgets*) or personification*: Two Brains are better than one*; *Introducing the new iPod family*.

From a pragmatic point of view, in addition to easily deduced implicatures, explicatures that require insignificant cognitive efforts on the part of the addressee of the advertisement prevail.

Thus, in comparison with the first stage of identity construction, the second stage, while maintaining an orientation towards a creative client, noticeably strengthens the identity component associated with the characteristics of the product itself. Apple's advertising campaign of this period primarily uses a lower hierarchy of human needs compared to that of the first period. The far peripheral zone of the safety need, associated with comfort, security and stability moves to the core values, while the previously core motivations, related to exploration and the search for new meanings becomes the near periphery. The Seeker archetype in the ITCI is less prominent than the Guardian archetype, and this trend persists during subsequent periods of identity construction.

Within the framework of the Optimal Innovation Hypothesis, the means constructing the ITCI can be interpreted as follows. Disregard for cooperative maxims remains an ostensive stimulus in the actualization of the "creativity" descriptor. Recognizability of the familiar in the novel is provided by intertextual allusions that allow such markers to be optimally innovative, placing them on the first scale of attractiveness. At the same time, among the slogans attributing human abilities to the product we identify those that simultaneously violate maxims and involve little or no familiarity if lacking visual support. Such pure innovative devices may rank lowest on the ITCI attractive scale. The majority of slogans, foregrounding product associated descriptors, have little novelty about them, but they are "quite pleasing" on account of their familiarity, and they are the most frequent.

The third conditionally distinguished period focuses mainly on the maintenance and strengthening of ITCI convenience and comfort, which, accordingly, determines the use of descriptors associated with the qualities of the advertised product. Convenience, accessibility and multifunctionality as the second level signification meanings associated with Apple products are iconically reproduced by linguistic and pragmatic means. Slogans are concise: from a syntactic point of view, these are mostly simple sentences (*We mean business*) or nominative sentences, indicating the advertised quality of goods (*The world's fastest computer; Movies, TV shows, games, and music*) as well as elliptical structures (*Mini. The next big thing*), which in a laconic form highlight the main characteristic of the advertising product.

From a pragmatic standpoint, slogans reveal compliance with the maxims of cooperation and correspond either to constative speech acts (*The world's fastest computer*) or the direct distance-reducing directives calling for the product to be tested (*Clip and go. Put some music on*) as shown in Table 3.

Table 3. ITCI present day descriptors in pragmatic and stylistic manifestations (2003–2006)

<table>
<tr><th colspan="3">Identity descriptors</th></tr>
<tr><td colspan="3">Multifunctionality, exclusiveness, convenience, economy, power (product-associated descriptors)</td></tr>
<tr><th>Slogans</th><th>Pragmatics</th><th>Stylistic devices</th></tr>
<tr><td>The fastest, most powerful iPhone yet.
Faster. Greener. Still mini.
The new MacBook Air. Better graphics. More storage. Yet still the world's thinnest notebook.
Redesigned in a very big way.
Twice as fast, for half the price.
There's an app for just about anything.
Meet the best iPods ever.
The world's most advanced operating system. Finely tuned.
A more immediate, intimate way to connect.
Our most personal device yet.
The most amazing iPhone yet .</td><td>1) adherence to maxims of information aimed at maximum informativity about the product advantages;
2) direct constative speech acts;
3) constative speech acts with expressive illocution, marked by quality intensifiers as the illocutionary force indicating device
4) distance-reducing directives</td><td>nominative sentences,
ellipsis,
positive-evaluative adjectives in comparative and superlative degrees</td></tr>
<tr><td colspan="3"><i>Innovativeness</i> (<i>product-customer associated descriptors:</i> metonymically relates the product and its owner).</td></tr>
<tr><td>Redesigned.
Reengineered,
Re-everythinged
Thinnovation
Bigger than bigger
Nano-chromatic
Completely Renanoed.
Apple reinvents the phone
This changes everything. Again</td><td>Disregard for cooperative maxims direct constatives</td><td>– tautology-based non sequitur;
– compounding-based occasionalisms;
– nominative sentences;
– ellipsis.</td></tr>
</table>

non-standard thinking; esthetic needs (customer-associated descriptors)		
The biggest thing to happen to iPhone since iPhone For the colorful Forward thinking The deeper you look – the more beautiful it becomes	Disregard for the maxims of quantity of information, based on tautology and reduction.	Nominative sentences Ellipsis aphoristic style
need for in-group affiliation		
Our most personal device yet The iPhone you have been waiting for The notebook for everyone. Now with more speed, power, and battery life A little video for everyone The all-in-one for everyone	positive politeness / involvement strategies; constative speech acts	Inclusive pronouns, simulated personal addressing *you*, Indefinite pronoun *everyone;* revitalization of the internal form of the word *personal* to implement the strategy of interiorization of the product.

In this period of ITCI formation, the descriptors appealing to the goods’ simplicity and convenience mainly define the customers’ motivational needs.

New at this stage is the ITCI descriptor “Professional”, which is based on the use of terms or an explicit indication of the purpose of the product for professional and business needs: *Sends other UNIX boxes to /dev/null; The 64-bit professional dream machine; We mean business.*

The “creativity” component of the client’s identity, which was in the motivational field core at the first stage of ITCI formation and which constituted the near periphery at the second stage, is moved to the far periphery, as it becomes more and more indefinite and ambiguous (iconically correlating with the disregard for the maxims of Manner-transparency of information). The stylistic and pragmatic means that highlight this descriptor are focused not so much on creativity as on play. The Seeker archetype in ITCI manifests in this respect elements of the Trickster / Jester archetype (experimenting with forms and meanings), based on (a) antonym-bound antithesis combined with a presupposition-based inference “Mini means iPad mini as a smaller version of the internet tablet iPad”, actualized by the one-word sentence "Mini": *Mini. The next big thing*; (b) aphorism-like laconic slogans of paradoxical semantics based on non sequitur (a stylistic device that combines semantically disconnected ideas): *Random is the new order; Enjoy uncertainty*; (c) a tautology (*Give chance a chance*) that triggers reflexivity and discursive implicature.

Such observations are confirmed by some studies, which identified that ludic stylistics of the "Trickster-Jester" role relies on explicit and implicit antithesis, oxymoron and other devices, which combine opposite, logically incompatible and mutually exclusive words or concepts. Pragmatically, "such devices correlate with the violation of Grice's maxim of manner (aimed at avoiding obscurity, ambiguities, or illogicalities), which triggers the discursive implicature – the inference of the meanings to restore semantic cohesion" (Kravchenko et al. 2020b: 190).

Within the scheme of the Optimal Innovation Hypothesis, the means shaping the ITCI at the third stage of Apple discourse development can be interpreted as follows. The slogans highlighting the descriptor "professional", and the product associated descriptors are at the second scale of attractiveness due to their recognizability by the target customers. The first scale of optimal innovative slogans remains "vacant", since ludic stylistics and experiments with forms and meaning associated with the "creativity" descriptor may remain unrecognized by the target client and, accordingly, moves to the lowest level on the ITCI attractiveness scale.

With reference to Maslow's hierarchy, this stage of ITCI formation, primarily satisfies the need for comfort, *convenience, economy* (second level motivations), the need to achieve success and career growth (fourth level), and, in part, the cognitive needs (fifth level). Predominant is the Keeper archetype and the Everyman archetype, as such in its subtype as the Networker (correlated with descriptor "Professional" in our study). Such a subtype of the Everyman archetype, in particular, is distinguished by works on branding strategies (Linabury 2018). The fifth level of needs is partially associated with the hybrid archetype of the Seeker-Jester, satisfying the needs of a certain group of consumers for curiosity and the search for new meanings.

3.3 The current stage of development of IT consumer identity

The current stage of development of ITCI is marked by the further strengthening of the descriptors related to IT goods' characteristics and aimed at the widest possible range of customers. In this regard, slogans are even more simplified in terms of both their structure and their semantics, highlighting one quality of the product, often associated with the features of its design, shape and size: *Faster. Greener. Still mini; The new MacBook Air. Better graphics. More storage. Yet still the world's thinnest notebook.* In addition to nominative sentences and ellipses (*Redesigned in a very big way*), which increase the expressiveness of slogans, a significant number of positive-evaluative adjectives appear in comparative and superlative degrees, which either intensify the attributive adjective to designate the advertised characteristic (*The fastest, most powerful iPhone yet; The best Windows app ever*) or simply state the superiority of Apple's product over the rest (*World's best*).

Marked by superlative form and / or the intensifier, the slogans in a form of constative speech acts can convey the expressive illocutionary force, since they satisfy the criterion of sincerity as the main felicity condition for expressives: *The most stunningly powerful iMac yet.*

At this stage of ITIC construction, the Apple discourse actualizes the new semantic descriptor "innovativeness", which is established both explicitly, by denotative semes "to invent" (*Apple reinvents the phone*) and "to change" (*This changes everything. Again*), and implicitly: *The first phone to beat the iPhone; This is only the beginning.* In the first case the slogans are, by their illocutionary force, constatives (describing the innovative characteristics), while in the second case they are indirect commissives containing the illocution of "promise of innovations". In this connection, the main felicity condition for commissives - namely, the ability of the promisor to fulfill his/ her promise – is restored in the context of the entire discourse of Apple.

At the present stage of ITIC, the first stage descriptor "in-group affiliation" is being actively updated again with involvement this time almost exclusively of the tactics of universalization instead of the tactics of "exclusiveness", which was widely used at the first stage of identity formation. The main means for marking the "group membership" descriptor (of Apple users) is the indefinite pronoun *everyone* to designate "every person", i.e. all people: *The new, faster MacBook Air. Everyone should have a notebook this advanced. And now everyone can; A little video for everyone; The all-in-one for everyone.*

The opposite tactic, the construction of an "in-group" of creative, unique users, is employed to a limited extent, and quite specifically in the form of experiments at the level of meanings and forms.

In particular, in terms of experimenting with meanings at the semantic-stylistic level, figures of inequality are widely used, including

(a) word-play: *For the colorful*: a denoted characteristic can be attributed both to a multicolored iPhone and to its owner – a bright non-standard person, who acquires this quality through the use of this product; *The biggest thing to happen to iPhone since iPhone*: combination of units denoting different but close notions of "significance" (of the event associated with the appearance of a new product) and a big size of display;

(b) tautology-based non sequitur (hinting at the big display of iPhone 6 Plus): *Bigger than bigger*;

(c) occasionalism, which, due to its internal form, simultaneously highlights two or more advertised characteristics: *Redesigned. Reengineered, Re-everythinged; Thinnovation; nano-chromatic; Completely Renanoed*.

Direct nomination of creativity as a quality of the target client is used in isolated cases. In addition, this characteristic is expressed in the form of a play on words, simultaneously attributing the quality of creativity to both the product and its owner: *Forward thinking; For the colorful.*

From the viewpoint of the Optimal Innovation Hypothesis, at the current stage of shaping Apple ITIC, the verbalizers of the "creativity" descriptor move to the first place on the attractiveness scale, being optimally innovative due to their simultaneous figurativeness and familiarity. However, in terms of their frequency they are significantly inferior to the means that highlight the "in-group affiliation" and "product-associated" descriptors, which lack figurativeness and novelty, but they are easily recognizable and, according to this criterion, are at the second stage of attractiveness.

Similar to the previous stage of the ITCI construction, the current advertising discourse of Apple, messaged in slogans *product-associated* descriptors "*convenience, comfort, economy*", primarily appeals to the consumer motivations correlating with the second level of safety and stability needs. The additional descriptor "superiority" (of Apple's product) partly relies on the fourth level esteem needs, including those associated with self-esteem, status and prestige. The descriptors "innovativeness" and "creativeness" metonymically connect the product and the customer-owner. As identity-forming descriptors, they are simultaneously associated with the sixth aesthetic level needs – appreciation and search for beauty, balance, form (product-associated facet) and with the fifth level cognitive needs of curiosity and exploration (customer-associated facet). The ITCI descriptor "in-group affiliation" is pertinent to the third level belongingness needs. Based on the identity descriptors, marked by linguistic and pragmatic means, the main archetypes at the present stage of the ITCT development include the Regular Guy (belonging), the Creator (innovation, imagination), the Keeper (stability and safety) and the Explorer (new experiences) as shown by Fig. 2.

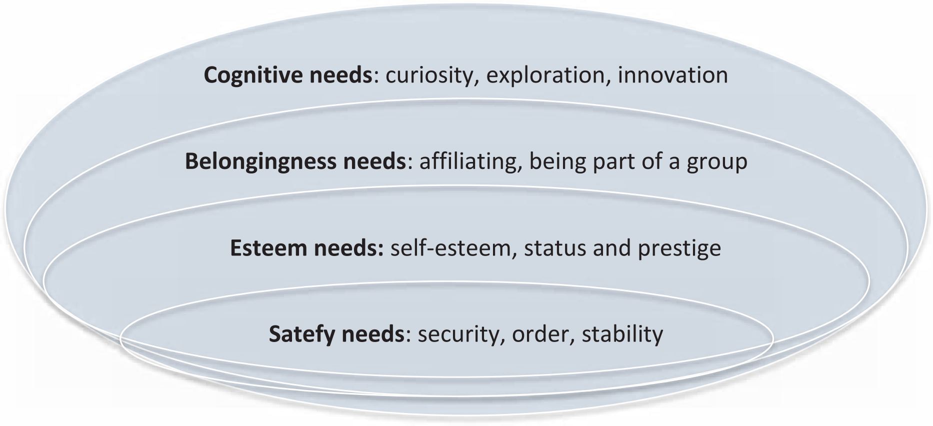

Cognitive needs: curiosity, exploration, innovation
Belongingness needs: affiliating, being part of a group
Esteem needs: self-esteem, status and prestige
Satefy needs: security, order, stability

Figure 2. Present stage of ITCI identity construction: Components-descriptors in the field structure of needs

4. Concluding remarks

This article examines the genesis and stages of formation of IT consumer identity based on the linguistic and pragmatic specifics of Apple's advertising slogans over the past 44 years.

The first stage of ITCI formation is based on the customer-associated descriptors "creativeness" and "need for in-group affiliation". Stylistically, "creativeness" is marked by oxymorons, allusions, puns and aposiopesis that iconically reproduce non-standard thinking, and pragmatically by disregard for cooperative maxims resulting in discursive implicatures. At the second and third stages, the descriptor is slightly modified by attributing the cognitive abilities to both the customer, through explicit performative "think", and the product, due to its personification, which in combination with allusion become a means of the extension of advertised properties of the thing on its owner.

The descriptor "need for in-group affiliation" is based on the opposing tactics of creativity-associated "specialness" and "universalization". The first vector, "creativity-based group identity", is clearly identified only at the first stage of ITCI formation. At the second and subsequent stages, this descriptor becomes less distinct, which is manifested by the use of ludic stylistics, paradoxes, and non sequiturs, and at the present stage by occasionalisms, which combine two advertised characteristics of a product by means of compounding. The Seeker archetype, corresponding to the "creativity" descriptor, hybridizes, acquiring elements of the Trickster-Jester.

The second vector, "universalization-based group identity", aimed at the maximum expansion of the "inner group" of Apple users, is marked by positive politeness strategies, indirect commissive speech acts, pronouns of inclusiveness (we, us, our) or instances of the simulated personal address *you,* included as a means of metonymic identification of the product with the personality of its owner. This vector is not revealed at the second and third stages of the ITCI formation, but is intensified and becomes one of the advertising priorities at the current stage, and is marked by the widespread use of the indefinite pronoun *everyone* in Apple slogans.

From the first stage of the ITCI formation to the present, the in-group expansion strategy also correlates with the product-related descriptors "multifunctionality", "simplicity", "availability", "innovativeness", and others. Such descriptors are pragmatically marked by adherence to the maxims of information, predominance of explicatures instead of implicatures in explicating the goods characteristics, constative speech acts, and the direct distance-reducing directives urging to test the product. Stylistically, slogans are becoming more and more simple both in their structure and in their semantics, and they are marked by ellipsis, partial parallelism with anaphoric subject, nominative case, one-part predicate sentence construction, highlighting the advertised operations that the product is capable of performing or facilitating. At the present stage the product-associated descriptor also relies on positive-evaluative adjectives in comparative and superlative degrees, which intensify the advertised characteristic and mark indirect expressive speech acts.

The values identified vary in terms of their frequency and duration of use in Apple advertising. A value such as creativity ("otherness", "specialness"), originated at the first stage of ITCI construction, remained discourse-forming for Apple advertising in the period 1997-2002. In 2003-2006 discourse, the same descriptor is conveyed mainly on the stylistic level of experiments with linguistic forms. In the subsequent stages of ITCI development, the customer-associated descriptor "creativity" transforms into the product-associated descriptor "innovativeness" (of Apple products), metonymically connecting the product and its owner-customer. In terms of frequency and duration of use, this descriptor is significantly "inferior" to such descriptors as "in-group identification" and product-associated descriptors. The latter are the most frequent and permanent for the APPLE discourse, their function as discourse-forming concepts has gradually increased at all stages of identity formation.

This indicates a steady trend towards a decrease in creativity aimed at maximum expansion of the group of clients by "betting" on a client with more ordinary motivational needs. A similar tendency is seen in the use of the descriptor "intragroup identification", which, starting from the second stage of ITCI formation, includes the tactics of universalization instead of the tactics of "exclusiveness", which were widely used in the first stage. In turn, tactics of "exclusiveness" are implemented in the form of intragroup identification (according to professional criteria), which embodies the descriptor "professional".

As regards the ITCI correspondence to the hierarchy of human motivations, the first stage of ITCI construction relies on the fifth level of the human needs hierarchy, that associated with exploration, a search for new meanings and the Explore archetype. The near peripheral zone close to the core corresponds to the third level belongingness need for being part of the consumer in-group, revealing the Regular Guy archetype, and far peripheral zone pertains to the second level safety need, which is related to the product-associated descriptors and the Keeper/ Guardian archetype. In contrast to the IT advertisement discourse of the first stage, that of the present-day foregrounds the product-associated values as the core components of the ITCI formation, while the fifth level (creativity-associated) cognitive needs move to the far periphery of the field of needs. Consequently, the customer-associated descriptor "creativity" transforms into the product-associated descriptor "innovativeness", and metonymically links the product and its owner-customer. The zone closest to the core is occupied by ITCT values such as the need for respect, while the need for in-group affiliation is a little further from the core and closer to the near periphery.

From the vantage of the Optimal Innovation Hypothesis, the means that actualize the same descriptors in different periods of Apple's discourse are placed in different ways on the scale of "attractiveness" for target customers, which affects their effectiveness in shaping identity. The means that reveal the descriptor "creativity", in different periods of identity formation, occupy 1 (are optimally innovative) or 3 scales (are purely innovative). Means that actualize the internal group identity of Apple customers, as well as descriptors associated with product properties, as a rule, occupy the second position on the scale of attractiveness, since, while they are not novel, they are well recognized.

In the future, it is planned to conduct an associative experiment with the involvement of Apple customers in order to confirm/clarify the identified descriptors, as well as to check the results related to the identified scales of optimal innovation and attractiveness of the slogans for target customers.

REFERENCES

Sources

List of Apple Inc. Slogans
https://annex.fandom.com/wiki/List_of_Apple_Inc._slogans#cite_note-H30B-2

List of 50+ Top Apple Brand Slogans
https://benextbrand.com/apple-brand-slogans/

Special studies

Ahuvia, A.
2005 "Beyond the extended self: Loved objects and consumers' identity narratives", *Journal of Consumer Research* 32 (1), 171-184.

Austin, J.
1970 *How to Do Things with Words*. Oxford: Clarendon Press.

Bach, K.
2010 "Impliciture vs explicature: What's the difference?" In: B. Soria – E. Romero (eds.) *Explicit Communication: Robyn Carston's Pragmatics.* London: Palgrave Macmillan, 126-137.
2012 "Saying, meaning, and implicating". In: K. Allan – K. Jaszczolt (eds.) *The Cambridge Handbook of Pragmatics*. Cambridge: Cambridge University Press, 47-68.

Bamberg, M. – A. De Fina – D. Schiffrin
2011 "Discourse and identity construction". In: S.J. Schwartz – K. Luyckx – V.L. Vignoles (eds.) *Handbook of Identity Theory and Research*. New York: Springer Science, 177-199.

Barthes, R.
1973 *Mythologies*. London: Paladin.

Bordron, J.-F.
2011 *L'iconicité et ses images. Études sémiotiques.* Paris: Presses Universitaires de France.

Brown, P. – S. Levinson
1987 *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Carston, R.
2004 "Relevance Theory and the saying/ implicating distinction". In: L. Horn – G. Ward (eds.) *The Handbook of Pragmatics*. Oxford: Blackwell, 633-656.

Cherrier, H. – J. Murray
2007 "Reflexive dispossession and the self: Constructing a processual theory of identity", *Consumption Markets & Culture* 10 (1), 1–29.

Dunn, K.C. – I.B. Neumann
2016 *Undertaking Discourse Analysis for Social Research*. Ann Arbor: University of Michigan Press.

Elliott, R.
2004 "Consumption as a symbolic vocabulary for the construction of identity". In: K. Ekström – H. Brembeck (eds.) *Elusive Consumption*. Oxford; New York: Berg Publishers, 129-140.

Erikson, E.H.
1980 *Identity and the Life Cycle*. New York; London: Norton.

Faber, M. – J. Mayer
2009 "Resonance to archetypes in media: There's some accounting for taste", *Journal of Research in Personality* 43 (3), 307-322.

Fairclough, N.
1992 *Discourse and Social Change*. Cambridge: Polity.
1995 *Critical Discourse Analysis.* London; New York: Longman.

Gabriel, Y. – T. Lang
2006 *The Unmanageable Consumer: Contemporary Consumption and its Fragmentation.* London: Sage Publications.

Grice, P.
1975 "Logic and conversation". In: P. Cole – J.L. Morgan (eds.) *Syntax and Semantics 3: Speech Acts*. New York: Academic Press, 41-58.
1989 *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Hammack, P.
2008 "Narrative and the cultural psychology of identity", *Personality and Social Psychology Review* 12 (3), 222-247.

Hodge, R.V. – G. Kress
1988 *Social Semiotics*. Cambridge: Polity Press.

Holzscheiter, A.
2014 "Between communicative interaction and structures of signification: Discourse theory and analysis in international relations", *International Studies Perspectives* 15 (2), 142-162.

Jung, C.G.
1969 *Archetypes and the Collective Unconscious.* In: G. Adler – R.F.C. Hull (eds.) *Collected Works of C.G. Jung.* Vol. 9 (Part I). Princeton: Princeton University Press.
1971 *Psychological Types.* In: G. Adler – R.F.C. Hull (eds.) *Collected Works of C.G. Jung.* Vol. 6. Princeton: Princeton University Press.

Kravchenko, N. – T. Pasternak
2018 "Claim for identity or personality face: The Oscar winners' dilemma", *Lege Artis. Language Yesterday, Today, Tomorrow* 3 (1), 142-178.

Kravchenko, N. – O. Zhykharieva
2020 "Sign-like pragmatic devices: Pro et contra", *Kalbų studijos / Studies about Languages* 36, 70-84.

Kravchenko, N. – M. Goltsova – I. Kryknitska
2020a "Politics as art: The creation of a successful political brand", *Journal of History, Culture and Art Research* 9 (3), 314-323.

Kravchenko, N. – V. Snitsar – V. Blidchenko-Naiko
2020b "Paradoxes of rap artists' role identity: Sage, Magician or Trickster?", *Cogito. Multidisciplinary research journal* XII (1), 179-195.

Kress, G.
2010 *Multimodality: A Social Semiotic Approach to Contemporary Communication*. London; New York: Routledge.

Leech, G.
2014 *The Pragmatics of Politeness.* Oxford: Oxford University Press.

Linabury, D.
2018 *What are Branding Archetypes and How Do They Work?* https://element5digital.com/what-are-branding-archetypes-and-how-do-they-work/, accessed November 2021

Lindenfeld, D.
2009 "Jungian archetypes and the discourse of history", *Rethinking History* 13 (2), 217-234.

Marion, G. – A. Nairn
2011 "'We make the shoes, you make the story'. Teenage girls' experiences of fashion: Bricolage, tactics and narrative identity", *Consumption Markets & Culture* 14 (1), 29-56.

Maslow, A.H.
1943 "A Theory of Human Motivation", *Psychological Review* 50 (4), 370-396.
1970a *Motivation and Personality*. New York: Harper & Row.
1970b *Religions, Values, and Peak Experiences*. New York: Penguin.

Mikkonen, I. – J. Moisander – A. Fuat Firat
2011 "Cynical identity projects as consumer resistance: The Scrooge as a social critic?", *Consumption Markets & Culture* 14 (1), 99-116.

Pearson, C.
2015 *Awakening the Heroes Within: Twelve Archetypes to Help Us Find Ourselves and Transform*. New York: HarperOne.

Potts, C.
2015 "Presupposition and implicature". In: S. Lappin – C. Fox (eds.) *The Handbook of Contemporary Semantic Theory*. Oxford: Wiley-Blackwell, 168-202.

Scollon, R. – S. Scollon
1983 "Face in interethnic communication". In: J. Richards – R. Schmidt (eds.) *Language and Communication*. London: Longman, 156-188.

Searle, J.
1969 *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Shadraconis, S.
2013 "Leaders and heroes: Modern day archetypes", *LUX: A Journal of Transdisciplinary Writing and Research* 3 (1), 1-13.

Shuval, N. – R. Giora
2005 "Beyond figurativeness: Optimal innovation and pleasure". In: S. Coulson – B. Lewandowska-Tomaszczyk (eds.) *The Literal and Nonliteral in Language and Thought*. Frankfurt am Main; New York: Peter Lang, 239-254.

Simpson, P.
2014 *Stylistics*. London: Routledge.

van Leeuwen, T.
2005 *Introducing Social Semiotics. An Introductory Textbook*. London: Routledge.

Wilson, D. – D. Sperber
2004 "Relevance Theory". In: L. Horn – G. Ward (eds.) *Handbook of Pragmatics*. Oxford: Blackwell, 607-632.

Zimmerman, D.H. – D.L. Wieder
1970 "Ethnomethodology and the problem of order: Comment on Denzin". In: J.D. Douglas (ed.) *Understanding Everyday Life: Toward the Reconstruction of Sociological Knowledge*. Chicago: Aldine, 285-298.

Address: Nataliia Kravchenko, Kyiv National Linguistic University, I.V. Korunets Department of English Philology and Translation, 73, Velyka Vasylkivska St., Kyiv, 03680, Ukraine.
ORCID code: https://orcid.org/0000-0002-4190-0924

Address: Olga Valigura, Kyiv National Linguistic University, Department of Oriental Philology, 73, Velyka Vasylkivska St., Kyiv, 03680, Ukraine.
ORCID code: https://orcid.org/0000-0003-0428-5421

Address: Vira Meleshchenko, Ternopil Volodymyr Hnatiuk National Pedagogical University, Department of Foreign Languages, 2, Maxyma Kryvonosa St., Ternopil, 46027, Ukraine.
ORCID code: https://orcid.org/0000-0002-3484-9905

Address: Liudmyla Chernii, Ternopil Volodymyr Hnatiuk National Pedagogical University, Department of Foreign Languages, 2, Maxyma Kryvonosa St., Ternopil, 46027, Ukraine.
ORCID code: https://orcid.org/0000-0002-4755-9536

Corpus-driven conversation analysis approach to mentor – mentee interactions in the context of practicum

Anna Bąk-Średnicka
Jan Kochanowski University of Kielce

ABSTRACT

Effective collaboration of a mentor–mentee type is built on nonhierarchical, non-directive, frequent, meaningful, (in)formal and compassionate relationships (e.g., Arshavskaya 2014; Izadinia 2015; Kim – Schallert 2011; Long et al. 2013; Mena et al. 2017; Moser et al. 2019). Such contact opens a space for constructive conversations that build the intellectual, knowledge and social capital of teacher candidates, their future pupils and mentors (Langdon et al. 2014). Contrarily, contact based on a highly hierarchical expert–novice type leads to a supervisory rather than a supportive relationship (Jones et al. 2016). The supervisory type negatively influences the challenging apprenticeship of observation (Lortie 1975) and the shaping of teacher candidates' teacher identity (Long et al. 2013; Palazzolo et al. 2019; Patrick 2013). This paper adopts a corpus-driven conversation analysis approach to nuances of *effective and good* mentor – mentee interactions during feedback sessions of student-teaching practica. The corpus consists of 109 utterances which were made by effective and experienced mentors and which are recorded in 11 transcript excerpts selected from three scientific articles. The utterances were used in this paper to develop a *framework of effective and good mentor communication*. This framework was built by assigning these 109 utterances to one of the three types of conversational frames outlined by Long et al. (2013), i.e., educative or supportive or evaluative, paired with one of three types of *eutoric* cues characterized as positive/good communication by Korwin-Piotrowska (2020), i.e., the human being/mentee or the topic or the conversation/dialogue. The findings show that there is a statistically significant difference between the frequency of utterances addressing the mentee and the topic in the educative frame and such frequency addressing the mentee and topic in the evaluative frame. In other words, in the educative frame utterances are topic-centered and in the evaluative frame they are mentee-centered. This framework can help in acquiring a better understanding of one's linguistic choices when interacting with others.

Keywords: conversation analysis, mentor-mentee interactions, practicum.

DOI: 10.25951/4825

1. Introduction

This paper utilizes a corpus-driven conversation analysis approach as it looks at mentor – mentee forms of communication by means of narratives enacted through face-to-face and online conversations. In particular, it refers to transcripts of mentor–preservice teacher conversations during feedback sessions of student-teaching practica. The point of reference is the thesis that such social interactions should reflect (family) relationships based on truth, sincerity and understanding, rejecting the policy of "smooth interpersonal functioning" (Korwin-Piotrowska 2020: 12, 55).

In this way, the paper addresses the neglected problem of reciprocal compassionate relationships in the context of pre-service teacher education in general and the quality of the practicum in particular (Kim – Schallert 2011; Long et al. 2013). Empirical findings point to the fact that effective collaboration of the mentor–mentee type is built on nonhierarchical, non-directive, frequent, meaningful, (in)formal and compassionate relationships (e.g., Arshavskaya 2014; Izadinia 2015; Kim – Schallert 2011; Long et al. 2013; Mena et al. 2017; Moser et al. 2019). Cultivating non-hierarchical relationships opens a space for productive conversations which further build the intellectual, knowledge and social capital of teacher candidates, their future pupils and mentors (Langdon et al. 2014; Long et al. 2013). Real human-human understanding via communication develops the parties' healthy personality (Nęcki 1996: 55). Contrarily, contact during feedback sessions based on a highly hierarchical expert – novice type leads to a supervisory type of relationship (Jones et al. 2016; Long et al. 2013). Such relationships negatively influence the shaping of teacher candidates' teacher identity (Long et al. 2013; Palazzolo et al. 2019; Patrick 2013; Soslau 2012). Most importantly, as stated by Kim – Schallert (2011: 1060; see also Dreer 2020: 677; Langdon et al. 2014: 93; Komorowska 2021), preservice teachers *may* project their (compassionate) relationships with mentors onto their future relationships with their own pupils.

Despite these tendencies, the relationship of the mentor–mentee type leaves much to be desired. For instance, as a rule of thumb (school) mentors and (school) administration tend to appraise mentees unrealistically, positively causing "an ideological meltdown" on the part of mentees (Smagorinsky et al. 2004: 22 quoted in Long et al. 2013: 181; Mena et al. 2017: 55). Also, making teacher candidates imitate "an assigned identity" rather than broaden their autonomy damages their confidence in developing their own unique (teacher) identity (Izadinia 2015: 5; Mena et al. 2017: 56-57).

In the end, an "[a]ddiction to praise can also reduce levels of motivation and autonomy" (Komorowska 2021: 38).

Accordingly, this paper examines attitudes hidden behind words and expressions reflecting specific conversational styles typical of *effective and good* post-lesson observation conferences. The inspiration for the introductory part of this paper is the Greek word *metaxú/"in-between"*. *Metaxú* characterizes human existence as being "between immanence and transcendence, the sacred and the profane, finitude and eternity, determinacy, individual and community, self and other, unity and plurality, the fear and the promise of plentitude…" (Duraj 2017: 5). In this instance, *metaxú* refers to a positive inter-human borderland, a *metaxú* sphere which is attainable through positive/good communication (Korwin-Piotrowska 2019, 2020). The rules of positive/good communication are discussed here, following Korwin-Piotrowska (2020), in the context of *eutoric*, a new branch of rhetoric, described as "mutual listening, empathy and a constructive dialogue" (Korwin-Piotrowska 2017: 20). In turn, the main part of this paper constitutes a corpus-driven conversation microanalysis of transcripts of utterances made by experienced and effective university student-teaching mentors. This microanalysis gives insights into patterns of mentor-mentee interactions approximating the positive inter-human borderland, the *metaxú* sphere. This paper offers a description of a *framework of effective and good mentor communication* as it emerged from the examined corpus.

It is hoped that this framework will help university mentors gain a better understanding of their own linguistic choices when interacting with mentees, and plan their future interactions appropriately.

2. The art of conversation

There are many complexities to human comprehension. We establish, maintain and develop contact with others by using (non)verbal communication. Verbal communication is, of course, possible by means of language, which is a tool used to express facts as well as opinions and intentions. It also represents a reality in human minds, and it reflects cognitive processes used in processing information. Language is represented by words and meanings which are related, but separate psychological entities. In fact,

> it is impossible to posit [a] one-to-one relationship between [a] linguistic form and meaning (or, [to] put it another way, between language form and function). The same linguistic, and inseparably,

> paralinguistic form can have [a] different meaning depending on the speaker (who is saying it) and the context (how the speaker perceives the situation and the relationships among the participants) (Tannen 2005: 10-11).

Many words can be used and understood both literally and figuratively. Indeed, "each word can have an unlimited number of meanings, especially when used metaphorically" (Nęcki 1996: 197). Apart from this, there are many additional complexities involved in readily available human-to-human comprehension, such as those related to situational contexts, human diversities and complex cross-relationships. While it is ultimately impossible to achieve a state of full mutual understanding, positive communication can generate a specific positive inter-human borderland (Korwin-Piotrowska 2019, 2020).

2.1 The practical aspects of mentor communication

For some, mentor – mentee conversations leave a lasting impression. This is especially possible when such conversations cross the inter-human borderland (Korwin-Piotrowska 2020: 151). Such extensive and real dialogic experience can be a critical moment in the (professional) lives of the mentor and mentee, comparable to *Kairos*, which builds an exceptional "bridge" which exists between two people for the rest of their lives (Korwin-Piotrowska 2020: 149, 155, 156). Undoubtedly, practicum mentors bear the prime responsibility for being leading light facilitators of school placements turned into communities of practice, of learning, and of support (Chaliès et al. 2010; Arshavskaya 2014; Krutka et al. 2014; Montecinos et al. 2015; Palazzolo et al. 2019).

In practice, mentors' roles take on a new aspect as they devote their time, space and readiness for the other party. Physically speaking, their body language should project acceptance, curiosity, engagement, openness, respect and trust. Psychologically speaking, mentors should cultivate the attitudes of mindfulness. Mindfulness is defined as a state of mind in which the attentional focus is in the present moment without judgment (Dekeyser et al. 2008; Kabat-Zinn 1990 quoted in Garner et al. 2018: 378). Linguistically speaking, mentors should be equipped with mediation skills. Their awareness of language ambiguity, its metaphorical nature and limitations can help them navigate a conversation expertly by stabilizing and monitoring it. Instances of these skills are reframing, renaming or recasting (Arshavskaya 2014: 136).

Thus, mentors′ mediation is characterized by "re-framing the teacher′s thinking about teaching toward a more expert way of thinking about teaching, re-naming the teacher′s conceptions of teaching through expert discourse, and promoting the teacher′s more expert understanding of teaching through the use of the pedagogical concepts of teaching" (Arshavskaya 2014: 136). For example, as stated by Arshavskaya (2014), the preservice teacher should strive to externalize the need to create a more compassionate relationship with pupils by stating that she wants them to feel that they are important and cared for so that the classes are special for them; the mentor should strive to reinforce this need of re-framing and re-naming this concept as a way of creating "a community of learners" who "invest more in the class than just getting the content" (Arshavskaya 2014: 133).

Mentors′ mediation also means carefully withdrawing and giving time and space to the other party rather than appropriating the right to be correct. For example, Vásquez reflects thus: "as I began to analyze these data, both the program director and I were shocked to discover how much talk we produced relative to the TAs [teaching assistants]. This realization led to significant changes in the ways we conducted future meetings" (2004: 42). Likewise, Mena et al. maintain that "only a small number of mentors manage to create an environment in which PSTs [preservice teachers] are encouraged to raise more general questions and to discuss their own concerns" (2017: 57, based on Harrison et al. 2005).

Also, mentors′ mediation includes differentiating between facts and opinions, as well as finding areas of partial understanding, similarity and agreement when it comes to diagnosing and solving (classroom) problems. The mentoring language used in a conversation should not be categorical, confusing, disorientating, humiliating, ironic or moralizing. Mentors should respect the mentees′ perceptions, interpretations and their unique judgements of (classroom) incidents and events. Positive communication is built on intellectual and emotional empathy. While the former enables the making of accurate predictions about behavior, the latter enables the awareness of what someone else is feeling and effectively communicating it (based on Korwin-Piotrowska 2020: 140-155 chapter III/5; 2019: 67).

2.1.1 The art of listening

The art of listening constitutes both reflective and dialogic listening. Reflective listening assumes a mentor – mentee intellectual and emotional empathic engagement in communication. Such listening is active, attentive, careful, close and intentional. Reactions include mimicry, gesture, posture, nodding, exclamations, commentary, and questions or silence. It is important to listen carefully all the way to the end of a sentence uttered, which helps in better understanding someone else's point. Other strategies enabling better understanding of what is "hidden between the lines" are paraphrases, clarifications and reflections (Suchańska 2007: 160).

Preservice teachers must feel safe to say whatever they think since "the more we can build trust, the more risk they'll take in their teaching and their learning, and the more they'll be willing to confront tough issues that will eventually shape their lives as teachers' (4-27-04, 3rd interview)" (Kim – Schallert 2011: 1065). Symptoms of a lack of reflective listening are disinterest, indifference and passiveness, as well as a willingness to make a rapid assessment and retort, trying to appropriate the content of what was said for one's own ends.

Dialogic listening moves beyond empathetic listening towards active intellectual and emotional engagement. Dialogic listening is possible by means of "sculpting mutual meaning" (Stewart – Thomas 2000: 234-256). For example, "if the majority of the discourse is centered on how the student teacher feels, recounting the lesson, and giving advice, as is typical of the *telling* style, then opportunities to discuss the complexities of learning how to teach and discovering the deep rationales behind decision-making are non-existent" (Soslau 2012: 777). Also, "while emotional support is an important aspect of the supervisor's role, by itself it does not allow for the development of a vision of ambitious instruction" (Long et al. 2013: 187). There are three reactions involved in this approach to listening which deepen mutual understanding towards working out new solutions. These reactions include asking for elaboration on a new issue, introducing metaphors, and using paraphrases as extensions of the interlocutor's ideas (Stewart – Thomas 2000: 247-251) (based on Korwin-Piotrowska 2020: 143-145 chapter III/5).

2.1.2 Three practical aspects of good mentor communication: The mentee, the topic, and the dialogue cues

Korwin-Piotrowska (2020) analyses forty-one linguistic choices with a view to achieving the inter-human borderland on the part of interlocutors. The choices are grouped in relation to (1) the human being, (2) the topic, and (3) the dialogue.

There are fifteen *eutoric* procedures to addressing a human being, including the following actions: *pro homine* & *pro personae bono*, encourage &

show interest, highlight the shared reality, use 1st person plural to develop a sense of community, suggest help/suggest sincerely, understand emotions, show respect for silence, support linguistically, support cognitively, show a positive attitude irrespective of disagreement, summarize mutual discrepancies, the act of *parrhesia*, a one-sided conversation, and introduce silence.

There are eighteen *eutoric* procedures related to a topic: treat a topic as a common cause, search for agreement, order and clarify issues, calibrate meanings, show subjectivity of judgements, accept the interlocutor's stances, find arguments for the interlocutor's thesis, delay (negative) answers, summarize in a subjective retrospective manner, ask for correction and verification, summarize partially with a question, hedge to maintain rapport, think alternatively, be empathetic with insights, highlight shared realities, suggest settling claims together, ask for questions and explanations, utilize paraphrases.

There are eight *eutoric* actions, which are metacommunicative in nature: provide comments showing appreciation of the dialogue, monitor and modify the dialogue, ask for clarification with reflection, self-correct, use politeness conventions, comment by means of politeness conventions in metacommunicative ways, signal topic changes as well as openings or closings of digressions, suggest undertaking decisions together (based on Korwin-Piotrowska 2020: 159-166).

2.2 Three practical aspects of effective mentor communication: Educative, supportive and evaluative frames

Long et al. (2013) analyze mentor-mentee interactions during post-observation meetings from the perspective of developing ambitious teaching. The vision of ambitious teaching has been a part of mathematics and science reform initiatives (Long et al. 2013: 180). Referring to a number of sources (Fennema – Romberg 1999; NAS 1996; NCTM 2000; NRC 2001, 2007; Windschitl et al. 2011), Long et al. suggest that

> ambitious mathematics and science teaching emphasizes a student-centred pedagogy that enables students to know and use mathematics and science knowledge, to reason mathematically and scientifically, to test models and provide evidence-based explanations, and to participate productively in mathematical and scientific practices and discourse (2013: 179-180).

Categorizing, or framing, conversational interactions between university mentors and mentees into educative, supportive and evaluative shows how language and conversational style influence the effectiveness of such interactions (Long et al. 2013: 181). The educative, supportive and evaluative frames were identified in supervisors' comments, explanations and suggestions and highlighted by using Jefferson Transcript Notation (e.g., Atkinson – Heritage 2006).

The educative frame is characterized as "a willingness to give and receive feedback, suggestions, and explanations" (Long et al. 2013: 193). An example of this frame is the following mentor's utterance: "*maybe just do sort of like a model of what they are going to do*" (Long et al. 2013: 184). The supportive frame in turn concentrates on emotional support, redirecting the conversation "away from critical comments"; an example of this frame is the following mentor's utterance: "*I understand, again, there's parameters and, you know, there's only so much you can do in certain situations*" (Long et al. 2013: 184). The evaluative frame is characterized by (extensive) discussions and repetitions "to emphasize importance of issue" rather than provide preservice teachers with the "opportunity to explore and develop a vision of ambitious instruction" (Long et al. 2013: 184). An example of this frame is the following mentor's opinion: "*I think you made very clear directions. That's one of the things I see that you do more and more as we go along. Your directions are clear and there's no confusion as I look around the room … so than you for being really top-notch in that regard*" (Long et al. 2013: 184).

3. Corpus-driven conversation analysis of mentor utterances in the context of feedback sessions

3.1 Research aims and questions

This part of the paper examines post-observation university mentors' utterances. The corpus of 11 excerpts and 109 utterances was accessed through three scientific articles by Arshavskaya (2014), Kim – Schallert (2011) and Long et al. (2013) which were selected from a set of 23 articles constituting datasets of a part of ongoing larger corpus-based research. The rationale behind this corpus selection is that it includes excerpts of effective and experienced mentor utterances. In these articles, the excerpts were examined to analyze affairs between university mentors and their mentees, such as mediation in dialogic exchanges (Arshavskaya 2014), caring relationships (Kim – Schallert 2011), and conversational frames (Long et al. 2013).

It is assumed that a close analysis of effective and experienced mentors' words and sentences/utterances can help develop a *framework of effective and good mentor communication*. In the course of this analysis there were four stages. First, the supervisors' turns were singled out from the conversations. Turns are defined here as uninterrupted flows of speech during turn-takings which occurred during those conversations. Next, the mentors' uninterrupted flows of talking were divided into utterances. If the filled pauses included such hesitations as 'uhm', 'well', 'you know' which referred to what had been said, they were counted as utterances. If these filled pauses were used to signal a new topic, they were counted as utterance boundaries. Likewise, the unfilled utterances, such as hesitations, were counted as utterance boundaries within a turn. Then, the 109 utterances were assigned to one of the three types of conversational frames, i.e., educative, or supportive, or evaluative (Long et al. 2013). Finally, the utterances were paired with one of the three types of *eutoric* cues characterizing good communication, i.e., the human being/mentee, or the topic, or the dialogue (Korwin-Piotrowska 2020). Finally, the utterances were counted and coded as numbers.

The following research question was formulated: What is a model mentor conversational style with reference to the educative frame?

3.2 Research analysis and results

Altogether, in the corpus of 11 excerpts there were conversations among eight preservice teachers and their six teacher educators at universities in the United States. As already stated, in transcript excerpts nos. 1, 2, 3, 4, 5, 6, 7 (Long et al. 2013), each mentor's uninterrupted utterances at each turn that took place were separated, counted and numbered.

Frames	Educative		Supportive		Evaluative
Sources	Arshavskaya 2014	Long et al. 2013	Long et al. 2013	Kim – Schallert 2011	Long et al. 2013
transcript	excerpt no 8	excerpts no 4, 5, 6, 7	excerpt no 3	excerpts no 9, 10, 11	excerpts no 1, 2
participants	1 mentor / 1 mentee	2 mentors / 2 mentees	1 mentor / 1 mentee	1 mentor / 3 mentees	1 mentor / 1 mentee
Contact	a blog (online)	face-to-face	face-to-face	TeachNet (online)	face-to-face

In transcript excerpts 8, 9, 10, 11 (Arshavskaya 2014; Kim – Schallert 2011) each mentor′s quoted utterances were selected and used in the sources. Furthermore, they were separated, counted and numbered in the order in which they appeared there. In this way, it was possible to keep a record of all mentors′ utterances in this corpus. In total, there are 11 transcript excerpts.

3.2.1 **Part 1** (see Table 1 in the Appendix)

This corpus illustrates the educative conversational frame. It is based on transcript excerpts nos. 4, 5, 6, 7, 8. Transcript excerpt **no 8** (Arshavskaya 2014) uncovers a teacher educator′s mediation with an MA TESL[1] preservice teacher via a dialogic blog as an assignment for the MA TESL teaching practicum. The teacher educator is a professor with extensive experience in ESL teaching, supervising and mentoring MA TESL preservice teachers (Arshavskaya 2014: 131).

Transcript excerpt **no 4** (Long et al. 2013) reveals how a university supervisor created partially educative experience in a face-to-face conversation with a mathematics preservice teacher. The university supervisor is a secondary mathematics supervisor with less than five years of teaching experience hired by a university to mentor student teachers during their practicum. The broader aim of the teacher education program was "to prepare student teachers to adopt ambitious teaching practices" (Long et al. 2013: 182). Transcript excerpts **nos. 5, 6, 7** (Long et al. 2013) present how a university supervisor provided a science preservice teacher with educative experience in a face-to-face conversation. The university supervisor is a secondary science supervisor with less than five years of teaching experience hired by a university to mentor student teachers during their practicum. The broader aim of the teacher education program was "to prepare student teachers to adopt ambitious teaching practices" (Long et al. 2013: 182).

These excerpts lead to productive conversations and discussions highlighting specific critical incidents and events and the asking of probing questions. Detailed analysis of the linguistic cues in these transcript excerpts allowed categorization of the utterances into the dominating educative experience (37 utterances nos. 1-37) as well as, single examples of utterances categorized into supportive (5 utterances nos. 38-42) and evaluative (3 utterances nos. 43-45) experience. Table 1 contains all 45 utterances.

[1] Master of Arts (MA) in Teaching English as a Second Language (TESL).

3.2.2 **Part 2** (see Table 2 in the Appendix)

This corpus illustrates the supportive conversational frame. It is based on transcript excerpts nos. 3, 9, 10, 11. Transcript excerpt **no. 3** (Long et al. 2013) shows the flow of a face-to-face conversation between a university supervisor and a biology preservice teacher. The university supervisor is a secondary science supervisor with over 20 years of teaching experience hired by a university to mentor student teachers during their practicum. The broader aim of the teacher education program was "to prepare student teachers to adopt ambitious teaching practices" (Long et al. 2013: 182).

Transcript excerpts **nos. 9, 10, 11** (Kim – Schallert 2011) contain a teacher educator's responses, via the online posting medium *TeachNet*, to three preservice literacy teachers' public reactions to their course readings. The teacher educator is "a professor of literacy, the lead instructor of the program preparing future primary teachers of literacy. A recognized figure both on campus and in the field, he was known as an excellent teacher educator" (Kim – Schallert 2011: 1061).

These excerpts are characterized mainly by emotional support with a limited amount of critical, deepened explanations of the matters in question. A detailed analysis of the linguistic cues in these transcript excerpts allowed the categorization of the utterances into the dominating supportive (24 utterances nos. 51-74), educative (5 utterances nos. 46-50) and evaluative (10 utterances nos. 75-84) experience. Table 2 contains all 39 utterances.

3.2.3 **Part 3** (see Table 3 in the Appendix)

This corpus illustrates the evaluative frame. It is based on transcript excerpts **nos. 1, 2** (Long et al. 2013). They provide an example of a face-to-face conversation between a university supervisor and a preservice mathematics teacher. The supervisor is a secondary mathematics supervisor with over 20 years of teaching experience hired by a university to mentor student teachers during their practicum. The broader aim of the teacher education program was "to prepare student teachers to adopt ambitious teaching practices" (Long et al. 2013: 182).

These excerpts are based on detailed descriptions of pupils' behavior with little or no input of how to refine the practicum. Detailed analysis allowed for a distinction between the dominating evaluative experience (19 utterances no 91-109), educative experience (4 utterances no 85-88) and supportive experience (2 utterances no 89-90). Table 3 contains all 25 utterances.

3.2.4 Sum up of parts 1-3

In total, there are 109 utterances in these 11 transcript excerpts, as shown below:

Frames	Educative	Supportive		Evaluative	
Sources	Arshavskaya 2014	Long et al. 2013	Long et al. 2013	Kim – Schallert 2011	Long et al. 2013
transcript	excerpt no 8	excerpts no 4, 5, 6, 7	excerpt no 3	excerpts no 9, 10, 11	excerpts no 1, 2
number of mentors’ utterances	14	31	13	26	25
total in frames	45		39		25
Total	109				

Frame Table	Educative	Supportive	Evaluative	Total
Table 1	37	5	3	45
Table 2	5	24	10	39
Table 3	4	2	19	25
Total	46	31	32	109
	109			

3.2.5 Part 4

In order to design *a framework of mentors’ effective and good communication* in this corpus, the 109 utterances, assigned to the three frames, were also assigned to 41 *eutoric* cues within the three groups: the mentee / the topic / the dialogue. The quantity of the 109 utterances as categorized into the three *eutoric* cues and frames are shown below:

Frames Eutoric Cues	Educative (46 utterances)	Supportive (31 utterances)	Evaluative (32 utterances)
the mentee	5	38, 39, 40, 41, 51, 53, 54, 55, 56, 57, 89, 90	43, 44, 45, 75, 76, 77, 78, 79, 80, 81, 82, 84, 100, 101, 102, 103, 104, 105, 109

the topic	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 46, 47, 48, 49, 50, 85, 86, 87, 88	42, 52, 58, 59, 60, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74	83, 91, 92, 93, 94, 95, 96, 97, 98, 99, 106, 107, 108
the dialogue		61	

3.3 Findings

In order to test whether there was a relationship between the *eutoric* cues and frame types, the χ^2 test was used. This test verified whether there was a statistically significant difference between the expected frequencies, i.e., as specified in the research question (What is a model mentor conversational style with reference to the educative frame?), and the observed frequencies in the contingency table. The contingency table included two variables associated with *eutoric* cues (category: the mentee, the topic, the dialogue) and conversational frame types (category: educative, supportive, evaluative). However, taking into account a single sentence in the category: dialogue (i.e., sentence 61), this category was excluded from this analysis. Consequently, the variable *eutoric* cues had two categories: the mentee and the topic. The χ^2 post-hoc analysis was based on adjusted standardized residuals. Additionally, the Cramér's V was used to assess the effect size.

The findings show that there is a statistically significant relationship between *eutoric* cues and conversational frames (see table 4).

Table 4. The relationship between the two types of eutoric cues and the three types of conversational frames in the corpus

Variables	*Eutoric* cues			
	the mentee		the topic	
	N	%	N	%
Frame type: educative	1	2.2%	45	97.8%
Frame type: supportive	12	40.0%	18	60.0%
Frame type: evaluative	19	59.4%	13	40.6%
$\chi^2_{(df = 2)} = 31.76$; $p < 0.001$; Cramér's V = 0.54				

Additionally, post-hoc analysis shows that there is a statistically significant difference between the frequency of the mentee category and the frequency of the topic category in frame types: educative (p < 0.001). Additionally, there is a statistically significant difference between the frequency of the mentee cues and frequency of the topic cues in the evaluative frame type: evaluative (p < 0.001). However, there is no difference between the frequency of the mentee cues and the frequency of the topic cues in the supportive frame type: supportive (p = 0.134).

To sum up, the model mentor conversational style emerging from this corpus shows that: (1) in the educative frame (characterizing effective interactions), the *eutoric* cues (characterizing good communication) focus on the topic (97.8%) rather than on the mentee or the dialogue; and (2) in the evaluative frame, the majority of *eutoric* cues address the mentee (60%) rather than the topic or the dialogue, as illustrated in Fig. 1.

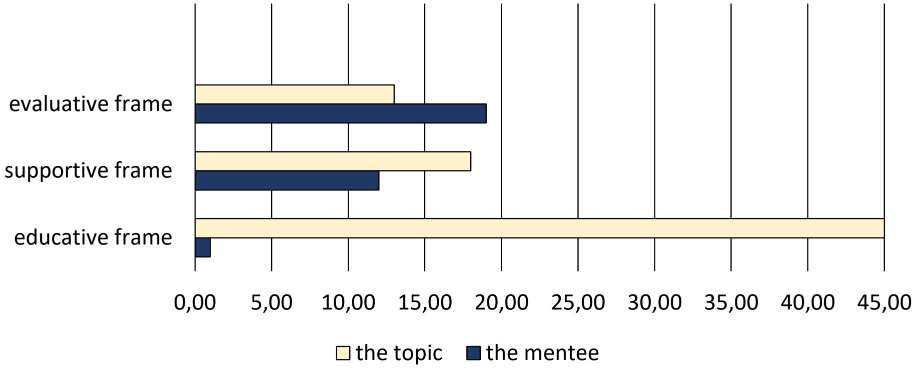


Fig. 1. The relationship between *eutoric* cues and conversational frames in the model mentor conversational style.

4. Conclusions

As outlined above, in preservice teacher education in general, and in the practicum context in particular, the matter of reciprocal compassionate relationships does not seem to be given the attention it deserves. Caring relationships between mentor and mentee positively contribute to future teachers' autonomy and readiness to take calculated risks. The mentor-mentee rapport enjoyed during the practicum positively influences future teaches' caring relationships with their pupils. It also enhances the development and maintenance of a healthy teacher identity.

The evidence scope of this article was limited to a corpus of mentors' exchanges in feedback sessions which were examined in three scientific articles, the aim of which was to pinpoint effective mentor-mentee interactions related to mediation in dialogic exchanges, building compassionate relationships and developing ambitious teaching though the educative conversational frame. The novelty here was the attempt to build a new framework of effective and excellent mentor conversational style as a result of analyzing this corpus. The framework was built on six categories of linguistic cues characterizing effective and clear communication. The conversational style emerging from this framework, which rests on the foundation of this corpus, shows that most of the utterances were educative in nature and concentrated on the topic.

To sum up, clear and effective communication in this academic setting depends on the sharing of similar views, the providing of reciprocal responses, and the cultivating of partnerships which benefit each interlocutor by means of instructive, mutual dialogue (Al-Mekhlafi 2010; Kim – Schallert 2011; Okraj 2017: 56).

REFERENCES

Sources

Arshavskaya, E.
2014 "Analyzing mediation in dialogic exchanges in a pre-service second language (L2) teacher practicum blog: A sociocultural perspective", *System* 45, 129-137.

Kim, M. – D.L. Schallert
2011 "Building caring relationships between a teacher and students in a teacher preparation program word-by-word, moment-by-moment", *Teaching and Teacher Education* 27, 1059-1067.

Long, J.J. – E.A. van Es – R.W. Black
2013 "Supervisor–student teacher interactions: The role of conversational frames in developing a vision of ambitious teaching", *Linguistics and Education* 24, 179-196.

Special studies

Al-Mekhlafi, A.A.
2010 "Student teachers' perceptions on the effectiveness of practicum and practicum supervisors", *Ajman University Network of Science and Technology Journal* 15 (2), 7-28.

Atkinson, J.M. – J. Heritage
2006 "Jefferson's transcript notation". In: A. Jaworski – N. Coupland (eds.) *The Discourse Reader* (2nd edn.). London: Routledge, 158-166.

Chaliés, S. et al.
2010 "Training preservice teachers rapidly: The need to articulate the training given by university supervisors and cooperating teachers", *Teaching and Teacher Education* 26, 767-774.

Dekeyser, M. et al.
2008 "Mindfulness skills and interpersonal behavior", *Personality and Individual Differences* 44, 1235-1245.

Dreer, B.
2020 "Towards a better understanding of psychological needs of student teachers during field experiences", *European Journal of Teacher Education* 43 (5), 676-694.

Duraj, J.
2017 "The hermeneutics of metaxy in the philosophy of Plato", *Studia Bobolanum* 4, 5-22.

Fennema, E. – T.A. Romberg (eds.)
1999 *Classrooms that Promote Mathematical Understanding*. Mahwah, NJ: Erlbaum.

Garner, P.W. – S.L. Bender – M. Fedor
2018 "Mindfulness-based SEL programming to increase preservice teachers' mindfulness and emotional competence", *Psychology in the Schools* 55, 377-390.

Harrison, J. – T. Lawson – A. Wortley
2005 "Mentoring the beginning teacher: Developing professional autonomy through critical reflection on practice", *Reflective Practice* 6, 419-441.

Izadinia, M.
2015 "A closer look at the role of mentor teachers in shaping preservice teachers' professional identity", *Teaching and Teacher Education* 52, 1-10.

Jones, M. et al.
2016 "Successful university-school partnerships: An interpretive framework to inform partnership practice", *Teaching and Teacher Education* 60, 108-120.

Kabat-Zinn, J.
1990 *Full Catastrophe Living: Using the Wisdom of your Body and Mind to Face Stress, Pain and Illness*. New York: Delacorte.

Komorowska, H.
2021 The role of attention in teacher education: A factor in the quality of European schooling", *Theory and Practice of Second Language Acquisition* 7 (1), 33-50.

Korwin-Piotrowska, D.
2017 "Konflikt z perspektywy erystyki i eutoryki", *Tematy i Konteksty* 7 (12), 20-43.

2019 "Wokół kategorii metaxú. Doświadczanie i wytwarzanie międzyludzkiego pogranicza za pomocą języka", *Tematy i konteksty* 9 (14), 54-69.

2020 *Eutoryka. Rzecz o dobrej (roz)mowie*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Krutka, D.G. et al.

2014 "Microblogging about teaching: Nurturing participatory cultures through collaborative online reflection with pre-service teachers", *Teacher and Teaching Education* 40, 83-93.

Langdon, F.J. et al.

2014 "A national survey of induction and mentoring: How it is perceived within communities of practice", *Teaching and Teacher Education* 44, 92-105.

Lortie, D.C.

1975 *Schoolteacher. A Sociological Study* (2nd edn.). Chicago: The University of Chicago Press.

Mena, J. – P. Hennissen – J. Loughran

2017 "Developing pre-service teachers' professional knowledge of teaching: The influence of mentoring", *Teaching and Teacher Education* 66, 47-59.

Montecinos, C. – H. Walker – F. Maldonado

2015 "School administrators and university practicum supervisors as boundary brokers for initial teacher education in Chile", *Teaching and Teacher Education* 49, 1-10.

Moser, M.K. et al.

2019 "A survey of world language cooperating teachers: Implications for teacher development", *Foreign Language Annals* 52, 873-890.

National Academy of Sciences (*NAS*)

1996 *National Science Education Standards*. Washington, DC: The National Academy Press.

National Council of Teachers of Mathematics (*NCTM*)

2000 *Principles and Standards for School Mathematics.* Reston, VA: National Council of Teachers of Mathematics.

National Research Council (*NRC*)

2001 *Adding it Up: Helping Children Learn Mathematics*. Washington, DC: The National Academy Press.

2007 *Taking Science to School: Learning and Teaching Science in Grades K-8*. Washington, DC: The National Academy Press.

Nęcki, Z.

1996 *Komunikacja międzyludzka*. Kraków: Wydawnictwo Profesjonalnej Szkoły Biznesu.

Okraj, Z.

2017 "Jedno zagadnienie, kilka wariantów dyskusji w szkole wyższej: specyfika, techniki, przykłady", *Studia z Teorii Wychowania* III (1/18), 55-67.

Palazzolo, A. – S. Shahbazi – G. Salinitri
2019 "Working towards change: The impact of mentor development on associate teachers and faculty advisors", *Interchange* 50, 321-337.

Patrick, R.
2013 "'Don't rock the boat': Conflicting mentor and pre-service teacher narratives of professional experience", *The Australian Educational Researcher* 40, 207-226.

Smagorinsky, P. et al.
2004 "Tensions in learning to teach: Accommodation and the development of a teaching identity", *Journal of Teacher Education* 55 (8), 8-24.

Soslau, E.
2012 "Opportunities to develop adaptive teaching expertise during supervisory conferences", *Teaching and Teacher Education* 28, 768-779.

Stewart, J. – M. Thomas
2000 "Słuchanie dialogiczne: lepienie wzajemnych znaczeń". In: J. Stewart (ed.) *Mosty zamiast murów. O komunikowaniu się między ludźmi*. Tłum. P. Kostyło. Warszawa: Wydawnictwo Naukowe PWN, 234-256.

Suchańska, A.
2007 *Rozmowa i obserwacja w diagnozie psychologicznej*. Warszawa: Wydawnictwa Akademickie i Profesjonalne.

Tannen, D.
2005 *Conversational Style*. *Analyzing Talk among Friends*. Oxford: Oxford University Press.

Vásquez, C.
2004 "'Very carefully managed': Advice and suggestions in post-observation meetings", *Linguistics and Education* 15, 33-58.

Windschitl, M. – J. Thompson – M. Braaten
2011 "Ambitious pedagogy by novice teachers: Who benefits from tool-supported collaborative inquiry into practice and why?", *Teachers College Record* 113 (7), 1311-1360.

APPENDIX

Table 1. 45 mentors’ utterances providing dominating **educative** (nos. 1-37) as well as supportive (nos. 38-42) and evaluative experience (nos. 43-45)

Educative frame	educative experience (suggestions for improvement)	(1) “… think of it as you are learning with the students… if you are not sure of the answer it is *OK* to say so but then to reason out loud why you think one particular answer is better than another, and let them reason out loud too, and together see if you can come up with the various ways in which you all understand the question/answer. (Blog entry, 3 February 2008)” (Arshavskaya 2014: 132-133).
		(2) “this *seems like* you have a good strategy. (Blog entry, 3 February 2008)” (Arshavskaya 2014: 133).
		(3) “’… *if* you make it [an ice-breaker activity] quick and fun those who come late will think they missed out on something and may want to come on time’ (Blog entry, 3 February 2008)”* (Arshavskaya 2014: 133).
		(4) “’Next class bring in something to eat (a box of donuts, or something small) use them as rewards for doing/saying something for the first 15 minutes and then quickly put them away – those coming in late will again think they missed something’ (Blog entry, 3 February 2008)” (Arshavskaya 2014: 133).
		(5) “… I appreciate your comments about caring for students – this is critical in creating a community of learners and in getting students to invest more in the class than just getting the content. (Blog entry, 6 February 2008)” (Arshavskaya 2014: 133).
		(6) “… *sounds like* giving them a bit of ownership over even a small aspect of your course (the topic of their presentations) has made a huge difference in terms of their engagement and motivation. (Blog entry, 8 March 2008)” (Arshavskaya 2014: 134).
		(7) “With all that instructional paraphrasing going on, and in particular for beginning-level students, making your recasts and expansions explicit/concrete is essential. You did a lot of this, by writing new words on the board, but *I noticed that* very few students wrote them down. You *might want* to either ask them to write down everything that you write on the board, or you can write it down during the break, and then either give the word/phrase list to the students or *perhaps* have them do something with the list (create a concept map, or tell a round-robin story using each of the words, etc.). This *just* makes your instructional paraphrasing more concrete, more permanent, so now they can refer back to these words/phrases in other contexts or for other assignments… (Blog entry, 10 April 2008)” (Arshavskaya 2014: 135).

Educative frame

educative experience (suggestions for improvement)

(8) “I *noticed* that you do a lot of instructional paraphrasing. This [instructional paraphrasing] is an excellent instructional strategy as it shows them [the students] that you are trying to build bridges between what they know and what you are trying to teach them. It also allows you to recast students’ contributions that may be difficult for the rest of the class to understand. (Blog entry, 10 April 2008)” (Arshavskaya 2014: 135).

(9) “… all this sort of situating makes your instructional goals clear, the content/skills of your lesson transparent. (Blog entry, 10 April 2008)” Arshavskaya 2014: 135).

(10) “… so, for example, you *might have begun* with something like, ‘Today we are going to continue our discussion of the professions, but today we are going to talk about the changing roles of men and women in different professions…’ (Blog entry, 10 April 2008)” (Arshavskaya 2014: 135).

(11) “This *may seem* trivial to you but it is essential that you and the students can walk away from any of your classes and articulate what they have learned and why.” (Blog entry, 16 April 2008)” (Arshavskaya 2014: 135).

֍֍֍

(12) “… *yeah. Okay*. So when you’re looking at … O*kay* … So what do you see when you look at this. … *right*. (Excerpt 4 lines 85, 87, 89, 91)” (Long et al. 2013: 189-190).

(13) “So, and that is exactly what the comment was that *maybe* we can try and go for a little more student talk and interaction…” (Excerpt 4 lines 93-94) (Long et al. 2013: 189-190).

(14) “… and there’s ways that you can do that. There’s lots of creative ways that you can have them *you know* do more exploration and just have it be less work on you – less teacher-driven…” (Excerpt 4 lines 96-97) (Long et al. 2013: 189-190).

(15) “… and more student-driven. So I just wrote down a couple of examples of maybe ideas like you could have them measuring the sides so they get out a ruler and they measure each side and then they…“ (Excerpt 4 lines 99-100) (Long et al. 2013: 189-190).

(16) “… kind of come up with some conclusions…“ (Excerpt 4 line 102) (Long et al. 2013: 189-190).

(17) “… like what did you notice about this triangle. Turn to your partner and talk to them about what you noticed. All sides are equal. Okay. Well what do you think would be a good name for this kind of triangle. And then what do you notice about this triangle … that kind of thing. (Excerpt 4 lines 104-106, 108)” (Long et al. 2013: 189-190).

֍֍֍

(18) “… if they were to work out this problem right here…” (Excerpt 5 lines 12-13) (Long et al. 2013: 191).

Educative frame	*educative experience (suggestions for improvement)*	(19) "… and be able to *you know* work it as if you were doing stoichiometry…" (Excerpt 5 lines 15) (Long et al. 2013: 191). (20) "… they would be able to come up well the reason why *you know* or the difference between normality and molarity and you can *sort of* explain it to them…" (Excerpt 5 lines 17-18) (Long et al. 2013: 191). (21) "… conceptually what means because *I think* students are still…" (Excerpt 5 lines 20) (Long et al. 2013: 191). (22) "… a little bit confused about the difference between normality and molarity… " (Excerpt 5 lines 22) (Long et al. 2013: 191). (23) "… I mean it's *really* simple lab which *you know would help* them…" (Excerpt 5 lines 24-25) (Long et al. 2013: 191). (24) "… but as you move forward with more complex concepts *I think it's helpful* to have that basic foundation." (Excerpt 5 lines 27) (Long et al. 2013: 191). (25) "*one more suggestion* that I have is *maybe* to visually show them what they will be doing during…" (Excerpt 6 line 131) (Long et al. 2013: 192). (26) "… the lab, so here I *would draw you know* even on the board or on your lab handout *you know*…" (Excerpt 6 line 133) (Long et al. 2013: 192). (27) "… where's the base where's the acid…" (Excerpt 6 line 135) (Long et al. 2013: 192). (28) "… right. and see here. Also you can even put the amount…" (Excerpt 6 line 137) (Long et al. 2013: 192). (29) "… and the known amount…" (Excerpt 6 line 139) (Long et al. 2013: 192). (30) "… and you're being asked for so that visually they can *sort of*…" (Excerpt 6 line 141) (Long et al. 2013: 192). (31) "… okay as I move through this activity when I do this this is exactly what I am doing…" (Excerpt 6 line 143) (Long et al. 2013: 192). (32) "right. and I think this would also help them to bridge this connection right here…" (Excerpt 6 line 145) (Long et al. 2013: 192). (33) "… as they are being asked to solve the unknown *okay*…" (Excerpt 6 line 147) (Long et al. 2013: 192). (34) "one more thing that *I thought may be* to *sort of* minimize the time that is spent on the procedural stuff getting ready *you know maybe* just do *sort of like* a model…" (Excerpt 7 lines 156-157) (Long et al. 2013: 192). (35) "… of what they are going to do. '*Okay*. First you're going to get you know your acid and you're going to put this in this beaker and you can sort of do it with them…" (Excerpt 7 lines 159-160) (Long et al. 2013: 192).

Educative frame	*educative experience*	*(suggestions for improvement)*	(36) "… model it with them and then you know you're going to get your titration and so on and so forth and here's how you're going *you know* add the number of base into how you are going to mix the different solutions. And here's how are you going to read it so that when…" (Excerpt 7 lines 162-164) (Long et al. 2013: 192). (37) "… they get into their groups they probably spent more than half the time trying to get the things ready I mean … yeah of course…" (Excerpt 7 lines 166-167) (Long et al. 2013: 192).
	supportive experience	no suggestions for improvement	(38) "'[W]hat you are experiencing is *pretty normal* for a novice teacher.' (Blog entry, 22 February 2008)" (Arshavskaya 2014: 134). (39) "This *may seem* trivial to you but…" (Blog entry, 16 April 2008) (Arshavskaya 2014: 135). ֍֍֍ (40) "okay when they get so when they before they were taking their notes at school, at home. Then what are they doing in class, just like working through practice problems" (Excerpt 4 lines 75-76) (Long et al. 2013: 189-190). ֍֍֍ (41) "… and I understand why you are trying to do this…" (Excerpt 5 lines 24-25) (Long et al. 2013: 191). ֍֍֍ (42) "I mean that's something you can easily go over a couple of times with them." (Excerpt 6 line 133) (Long et al. 2013: 192).
	evaluative experience		(43) "It was a pleasure to watch you teach yesterday. It is obvious to me that you have all of the management/procedural strategies and techniques of an experienced teacher. (Blog entry, 10 April 2008)" (Arshavskaya 2014: 135). ֍֍֍ (44) "… always thinking academically. Academic language. So, that's a really good tool for that, and then also a couple of times when you were asking the kids questions you did a good job of having them explain why. So you said why is that when they were responding and then I *really* liked the closure with the sorting game, so that that was *really* good. So as for next time. So think. *Okay*. So let's look at the form here and just see *like* what you see when you're *like* looking at that in terms of what's going on (Excerpt 4 lines 56-61)" (Long et al. 2013: 189-190). ֍֍֍ (45) *"yeah I think* they're still not clear on this. The purpose of the lab but…" (Excerpt 6 line 149) (Long et al. 2013: 192).

Table 2. 39 mentors' utterances providing educative (nos. 46-50) as well as dominating **supportive** (nos. 51-74) and evaluative (nos. 75-84) experience

Supportive frame	educative experience (suggestions for improvement)		(46) "She was one of only two people, *I think*, who actually responded to this reading. I just love the way, I love that paragraph … the idea that in reader response, we're really talking about a relationship of a person with the text or the author is a new way of thinking for them, and she *really* put it together nicely, *I thought*. (4-27-04, 3rd interview)" (Kim & Schallert 2011: 1064). (47) "I'm glad you took it on … she's *really* a wonderful person … you would like her a lot … but that's *probably* not relevant … *right*? (Paul Jones/3-29-04)" (Kim & Schallert 2011: 1065). (48)"but never perfect … always a limitation … careful … 'the most cited research journal in education.' Do you really think it *would* get in this journal if it *were* flawed…? *The fact is* that there has been plenty of opinion but no data on this… (Paul Jones/3-29-04)" (Kim & Schallert 2011: 1065). (49) "Yikes … I am going to send this review to her … (Paul Jones/3-29-04)" (Kim & Schallert 2011: 1065). ֍֍֍ (50) "And I just had a question here. In the question it says 'three theories' and then I heard you using the word 'hypotheses'. Do they use them interchangeably or … " (Excerpt 3 lines 34-35) (Long et al. 2013: 188).
	supportive experience	no suggestions for improvement	(51) "There's no question mark on Michelle. She'll be great. She *just* needs more space for her personality to come out a bit" (Kim & Schallert 2011: 1063). (52) "hold on to this … despite what you might hear, this is the assumption you must make… (Paul Jones/3-23-04)" (Kim & Schaller 2011: 1063). (53) "question everything … I love the notion of being critical but not cynical … ask questions because we can learn. (Paul Jones/3-29-04)" (Kim & Schallert 2011: 1063). (54) "I watch for this because this is new … *I mean*, if you look back through her comments, this is the first time she's taken a stance, *I think*, and this isn't very strong but at least it's a start. (4-15-04, 2nd interview)" (Kim & Schallert 2011: 1063). (55) "She is having great tutoring experience with just a wonderful kid … It's just been *really, really* positive … She's the one I know the least … But again, I don't worry because I've got another year with them. It often happens that there are some kids that *sort of* stay in the background during the first semester and as they begin to get into a classroom situation, they come out *a little bit* more. (4-15-04, 2nd interview)" (Kim & Schallert 2011: 1063).

Supportive frame

supportive experience

no suggestions for improvement

(56) "There isn't a level of understanding or trust yet of her for me that she can open herself up yet. It'll happen. It's just going to take a little bit more time" (4-15-04, 2nd interview)" (Kim & Schallert 2011: 1065).

(57) "she must feel safe to say whatever she thought in TeachNet" (4-27-04, 3rd interview)" (Kim & Schallert 2011: 1065).

(58) "[T]he more we can build trust, the more risk they'll take in their teaching and their learning, and the more they'll be willing to confront tough issues that will eventually shape their lives as teachers' (4-27-04, 3rd interview)" (Kim & Schallert 2011: 1065).

(59) "*HA!!!* And stay out of trouble as well? No fear … I don't know this author. Blast away (Paul Jones/4-07-04)" (Kim & Schallert 2011: 1065).

(60) "We've got to learn to respond to each other and that will take a little time" (6-17-04, 4th interview)" (Kim & Schallert 2011: 1066).

(61) "I have to be *really* careful about how I respond to her' (6-17-04, 4th interview)" (Kim & Schallert 2011: 1066).

(62) "I have to be *really, really* careful. I'm not as circumspect as I should be in responding to certain people. We've got to learn to respond to each other and that will take a little time. (6-17-04, 4th interview)" (Kim & Schallert 2011: 1066).

֍֍֍

(63) "we were discussing the way lesson was broken up so the first part. You did warm-up and you had an agenda on the screen for them and they revisited some materials you'd done before. And looking at the three theories of how life on earth originated. *Okay*. And they you gave them a few minutes while you went around and stamped the homework…" (Excerpt 3 lines 1-4) (Long et al. 2013: 188).

(64) "So this homework is like something that they already have printed out ahead of time or just something you assign each day for … practice … last class…" (Excerpt 3 lines 6-7, 9, 11) (Long et al. 2013: 188).

(65) *"Okay. All right*. So, you stamped that and then how will you check for completion on that. *Just* you basically as you go around and stamp you're just looking for completion…" (Excerpt 3 lines 13-14) (Long et al. 2013: 188).

(66) *"Okay. All right* and then you used a random call method to choose students to give their answers that they'd written for the three theories and answered hydrothermal vents and then you asked some other additional questions that the other students answered. And then Mark said meteorites…" (Excerpt 3 lines 16-18) (Long et al. 2013: 188).

(67) "and again, you elaborated ad asked some more questions. And then Steven said pass…" (Excerpt 3 line 20) (Long et al. 2013: 188).

Supportive frame	*supportive experience*	*no suggestions for improvement*	(68) “does Steven pass very often…” (Excerpt 3 line 22) (Long et al. 2013: 188). (69) “… but Vicky volunteered to answer…“ (Excerpt 3 line 24) (Long et al. 2013: 188). (70) “and he said that large bacteria is that what she said…“ (Excerpt 3 line 26) (Long et al. 2013: 188). (71) “… found underneath the surface of the earth.” (Excerpt 3 line 28) (Long et al. 2013: 188). (72) “… and then you elaborated on that as well. *Okay* and a few more students had some input there…” (Excerpt 3 line 30) (Long et al. 2013: 188). (73) *“All right*. So then you told them to put the warm up away and all that took about 15 minutes or so. …(Excerpt 3 line 32) (Long et al. 2013: 188). (74) … with hypotheses *okay I was just curious*” (Excerpt 3 line 38) (Long et al. 2013: 188)”
	evaluative experience		(75) “I am so excited about this! (Paul/Jones/3-29-04)” (Kim & Schallert 2011: 1063). (76) “’She [Michelle] is wonderful. She’s strong. She’s thoughtful. She’s deep. She’s genuine. I mean, she’s a dream child’ (6-17-04, 4th interview)” (Kim & Schallert 2011: 1063). (77) “Nancy’s a hoot. She’s even more than what I’d hoped for in terms of being lively, hardworking, generous, just part of what you want in a cohort. She’s everybody’s buddy. (4-15-04, 2nd interview)” (Kim & Schallert 2011: 1064). (78) “*ok* … you are a convert! … Perfect. (Paul Jones/4-14-04)” (Kim & Schallert 2011: 1064). (79) “With Nancy, I know I can go at it directly and say, ‘change this, change that’ (4th interview)” (Kim & Schallert 2011: 1064). (80) “Nancy. Always good. You just look forward to it. I mean, it’s one of those you look forward to opening because you’re going to smile” (6-17-04, 4th interview)” (Kim & Schallert 2011: 1064). (81) “What a response!!! *Wow* … better than the article … so many of your qualities are revealed through this response” (Kim & Schallert 2011: 1065). (82) “What’s *really* good about her responses is that she always speaks from the heart. (4-27-04, 3rd interview)” (Kim & Schallert 2011: 1065). (83) “In many ways Goldy reminds me of [the reading’s author]. They’re very similar in terms of how, there’s this passionate way of how they view everything, and it was like an odd juxtaposition of two personalities” (4-27-04, 3rd interview)” (Kim & Schallert 2011: 1065). (84) “’She’s delightful. She’s honest. She thinks. She challenges. Goldy’s wonderful’ … (6-17-04, 4th interview)” (Kim & Schallert 2011: 1065, 1066).

Table 3. 25 mentor's utterances providing educative (nos. 85-88), supportive (nos. 89-90) as well as dominating **evaluative** (nos. 91-109) experience

Evaluative frame	educative experience(suggestions for improvement / reflection)		(85) "what would be your impression of the class today as far as were they engaged were there a bunch of kids that were absolutely tuned out not involved what would be your overall impression…" (Excerpt 1 lines 84-85) (Long et al. 2013: 185). (86) "… they were doing some considerable work" (Excerpt 1 line 116) (Long et al. 2013: 186). ֍֍֍ (87) "if it's just oral learning it's not going to work for these kids…" (Excerpt 2 line 239) (Long et al. 2013: 187). (88) "I thought you used some very high level questioning responding skills in here and examples of that are here's the big question how did you find it." (Excerpt 2 line 265) (Long et al. 2013: 187).
	supportive experience	(no suggestions for improvement)	(89) "contrary to what you might have felt – …" (Excerpt 1 line 116) (Long et al. 2013: 186). ֍֍֍ (90) "or maybe even less than you expected…" (Excerpt 2 line 259) (Long et al. 2013: 187).
	evaluative experience		(91) "that's interesting that you brought up *yeah* Michael. I didn't have his name. I just said you direct a boy to a text. Boy being Brandon." (Excerpt 1 lines 96-97) (Long et al. 2013: 186). (92) "Now that I know it's Michael and you called it something like a weird red thing and he got it he went over and picked up a textbook." (Excerpt 1 lines 99-100) (Long et al. 2013: 186). (93) "… and came back to his chair and did look in there as a source of info and did get involved in that. Then what I saw with him was he asked one of the boys up here, who is obviously probably one of the really bright dudes, *or something."* (Excerpt 1 lines 102-104) (Long et al. 2013: 186). (94) "… for help and he wanted the boy to come back here and the boy wanted to stay up there and I wanted to say come on Brandon. Go up there it's all right." (Excerpt 1 lines 106-107) (Long et al. 2013: 186). (95) "… there is an empty chair but that is not my place. I am supposed to be a fly on the wall here, so I didn't but then what I noticed was that he wasn't maybe as you suspected he wasn't totally not involved in my estimation anyway. He where [researcher] sitting he asked that boy right in front of him for some help…" (Excerpt 1 lines 109-112) (Long et al. 2013: 186).

Evaluative frame	*evaluative experience*	*(no suggestions for improvement)*	(96) "… and there was some dialogue back and forth which I was pleased. [student] it seemed like the one on her left there. Right here where I'm sitting did have some things going on and there was some and my listening to them was again…" (Excerpt 1 lines 114-116) (Long et al. 2013: 185-186). ֍֍֍ (97) "at least as half the class…" (Excerpt 2 line 231) (Long et al. 2013: 187). (98) have a background as second language learners and I think what I am seeing is that you make provisions for that I know in your lesson planning you do and in your actual teaching with the visual things you put up on screen…" (Excerpt 2 lines 234-236) (Long et al. 2013: 187). (99) and the ways that you represent so that *because you know* and we've talked about it at great length *about you know…"* (Excerpt 2 line 238-239) (Long et al. 2013: 187). (100) "… and that's not what you are doing so thank you for that. I think you made very clear directions. That's one of the things I see that you do more and more as we go along. Your directions are clear and there's no confusion as I look around the room there's not a lot of kids looking rolling their eyes looking in – having to ask for repeat instructions…" (Excerpt 2 lines 241-244) (Long et al. 2013: 187). (101) "… so thank you for being really top-notch in that regard. You use some choral responses which is fine. I also know that you often go to the cards and do. I don't think it was appropriate today that you needed to do that…" (Excerpt 2 lines 246-248) (Long et al. 2013: 187). (102) "… but you often do use the equity methods." (line 250) (Long et al. 2013: 187). (103) "Today I think you were using more choral responses not yelling out. I think there's a differentiation there. When you wanted a choral response you got it. When you wanted to call of someone you got that…" (Excerpt 2 lines 252-253) (Long et al. 2013: 187). (104) "so I see you as clearly in charge in that way…" (Excerpt 2 line 255) (Long et al. 2013: 187). (105) "… rather than *you know* having the kids in charge or some other rather loosey goosey format. It is pretty clear cut. I thought that there was some very powerful pieces that you did during he warm-up which by the way seemingly timed-out bout right…" (Excerpt 2 lines 257-259) (Long et al. 2013: 187). (106) "… but the other obviously took longer which is fine *you know* the body of the lesson which I would if I had to vote I'd rather

Evaluative frame	*evaluative experience*	*(no suggestions for improvement)*	see the body of the lesson take longer than the warm-up…” (Excerpt 2 lines 261-262) (Long et al. 2013: 187). (107) “… get way out of line or something and it didn’t. … You were getting some answers about distances and things. Some specific *you know*. Almost I’ll call them procedural kind of answers but then you went for what I call conceptual learning you were asking for them to clarify and tell you what was going on…“ (Excerpt 2 lines 264-268) (Long et al. 2013: 187). (108) “… and you were and anybody do anything different. Those kind of responses to me indicate your growing ability and incredibly good ability to go beyond the superficial questioning and that’s what it’s about in teaching and a lot of us don’t get there for years…“ (Excerpt 2 lines 270-272) (Long et al. 2013: 187). (109) “… you are getting there and that’s great” (Excerpt 2 line 274) (Long et al. 2013: 187).

Address: Anna Bąk-Środnicka, Uniwersytet Jana Kochanowskiego w Kielcach, Instytut Literaturoznawstwa i Językoznawstwa, ul. Uniwersytecka 17, 25-406 Kielce, Poland.
ORCID code: https://orcid.org/0000-0001-8932-659X

When *possible* does not always mean "possible": Evaluative patterns of newsworthiness in letters to the editor

Isabella Martini

University of Florence

ABSTRACT

The relevance of letters to the editor (LTE) calls for more research on the linguistic construction of their newsworthiness, particularly when letters are used to foster debate on controversial issues. The connection between newsworthiness and the language of evaluation has been studied quite extensively using corpora (Hunston 2011; Bednarek – Caple 2019). However, limited research has been performed on corpora of LTE to investigate their linguistic features (Pounds 2006; Romova – Hetet 2012). A small but significant corpus of LTE of *The Times* written between 1914 and 1926 on the Armenian question was selected to investigate their evaluative patterns of newsworthiness. Word frequency, collocational patterns, clusters of evaluative lexico-grammatical items and their semantic connotation were examined, also in relation to elements of the grammar of modality, with a specific focus on the evaluative adjective "possible" in its attributive and predicative uses. Understanding the linguistic strategies that contributed to keep alive the debate on those events provides further insights into the acknowledgment of the Armenian genocide.

Keywords: Letters to the Editor; Corpus Linguistics; News Discourse; Evaluation; Historical English.

1. Introduction

A considerable amount of textual material has been published on the Armenian genocide, both as first hands accounts, in form of diaries or interviews books, and news articles and letters to the editor (LTE) of international newspapers, such as *The Times*. As Peltekian (2013) remarks

DOI: 10.25951/4826

in the introduction to her collection of news articles and letters to LTE collected from the British press, the massacres of the Armenians were documented by war correspondents on a regular basis and kept alive in the section dedicated to the letters to editor of newspapers such as *The Guardian* and *The Times*.

As a form of mediated news discourse, LTE are ascribable to a genre with specific textual features that are worth investigating through a corpus-driven linguistic approach (Tognini-Bonelli 2001; Sinclair 1996, 2004). The Letters to Editor on the Armenian Question (LEAQ) small corpus of 186 LTE published between 1914 and 1926 was built from the online archive of *The Times and The Sunday Times* (https://www.thetimes.co.uk/archive/), which hosts the complete collection of the articles published between 1785 and 1985. Letters were selected using the key words *Armenia* and *Armenian* (the latter includes also mentions of *Armenians*) and has been analysed to collect corpus-driven quantitative and qualitative evidence on the evaluative language (Hunston – Thompson 2000) and the semantic prosody or evaluative connotational meaning (Sinclair 2003; Morley – Partington 2009) used to construct the newsworthiness (Bednarek 2006, 2010; Bednarek – Caple 2017, 2019) of the events connected to the Armenian situation in those years.

The question of the Armenian "relocation", i.e., the outbreak of violence on the Armenian residing in Anatolia between 1915 and 1918 (Elayyadi 2017), came back into international news when the war in the Nagorno-Karabakh area broke out in 2020, and the Armenian residents were forced to leave the area. While the Armenian genocide is being given more and more international recognition (Astourian 1990, Aybak 2016), Turkey denies responsibility for the Armenian genocide, ascribing the deportation and the massacre of around 1,5 million Armenians to the natural occurring events of the concurring First World War (Alayrian 2018).

LTE mentioning the Armenian question in the 20th century have not been analysed using a linguistic approach yet; building a corpus of letters mentioning the Armenian question and performing analyses on its linguistic features provides further research materials to answer two research questions:

- How was the reading public influenced in their perception of what was to be identified as the first genocide of the 20th century? And how was the language of evaluation used to construe news items as newsworthy and relevant in order to do so?

- With the Armenians striving to have the memory of the genocide recognised and kept alive, which linguistic strategies, if any, might have contributed to its general oblivion?

After a brief introduction of the historical context, the paper outlines the theoretical and methodological framework applied to the corpus of LTE. Then the construction of the corpus will be explained, and the evaluative patterns of newsworthiness discussed. The analysis will focus on the most recurrent evaluative adjective *possible* and on its attributive and predicative occurrences (Biber et al. 2007) in collocational patterns and clusters (Hunston 2002). Concluding remarks on further research paths are provided at the end of the paper.

2. The Armenian genocide. Some contextual information

On 24th April 1915, notable personalities of the Armenian minority living under the Ottoman rule were murdered in Istanbul; simultaneously the order was issued to kill Armenian men throughout the Ottoman empire, and to force the remaining members of the Armenian families to leave their homes and villages and march towards the Syrian desert. Civilians were forced to walk through villages with no one allowed to help them, exposed to constant brutality in a mass deportation that immediately caused international concern thanks to war correspondents and to high profile Armenians and international citizens living in those areas, who informed the international community of the atrocities perpetrated on the Armenians on a regular basis (Alayarian 2018).

Despite articles and LTE continuously mentioning the killings and the conditions of the deported Armenians in the international press, the Turkish government denies responsibility for the genocide (Chabot et al. 2016; Elayyadi 2017; Mamali et al. 2018). The Young Turks achieved a preeminent position in the years immediately preceding World War I and contributed to ignite the nationalist trend of the majority of the inhabitants of the empire. This led to the desire to "turkify" the Empire by removing the Christian minorities living within its borders – Armenians and Greeks, mostly – and to the wholesale massacre of civilians belonging to these minorities (Alayarian 2018; Mayersen 2016).

As outlined in the next section, news coverage of the events contributed to remind the international community of the crimes perpetrated by the

Ottoman government; LTE were used to keep the debate ongoing and to provide a space for high profile contributors to keep their memory alive.

3. Letters to the editor. Genre and corpus linguistics

LTE have achieved the status of a genre of its own within media discourse studies because of their peculiar features (Cavanagh 2019). Started as a space to share hard news, they later became a privileged space to share opinions and to make one's opinion known to the public. LTE ensured their writers not only visibility, but also recognition as a voice worth listening to (Hobbs 2019). This particularly happened in broadsheet newspapers such as *The Times*; high profile contributors could either respond to a specific matter or initiate a new conversation on a topic selected for its public significance (Brownlees et al. 2010).

LTE are usually written by members of the reading public of a newspaper, and their main aim is to communicate the writer's views'. Published letters sometimes undergo an editorial process that alters the authorial voice, thus creating a mediated news discourse suitable to reinforce the editorial line of the newspapers where they are featured, and to guide the reading public towards a specific reaction, thus generating a guided debate that mirrors the contents published in the newspaper (Richardson – Franklin 2004; Pounds 2006). LTE published on broadsheet newspapers, however, serve a wider and more strategic aim, as usually those are newspapers where matters of international politics are discussed by their actual protagonists, and where the debate in the empowered space dedicated to the LTE makes public what is otherwise privately discussed (Cavanagh 2019).

Despite their relevance to the construction and the performance of cultural citizenship (Cavanagh 2019), as well as their role in the construction of the media discourse in newspapers through the centuries (Hobbs 2019), and their availability in digitised formats, LTE have not been frequently analysed through a corpus linguistic approach, with the exceptions of Chovanec (2012), Romova and Hetet (2012) and Pounds (2005, 2006). Among these, Pounds (2006) examined the language of evaluation in the LTE in different cultural contexts (Italian and British), and her analysis provided insightful data on LTE as a tool of democratic participation and public engagement that contributed to the study conducted in this paper.

The rationale behind the creation of the corpus and the methodological framework of the analysis will be explained in the next section, with a specific focus on the parameters of the news discourse value analysis used to examine the evaluative function of adjectives in the corpus.

4. The language of evaluation applied to the LEAQ corpus

The language of evaluation has been the object of extensive linguistic research. A seminal formulation of the concept was made by Hunston and Thompson (2000); according to them, evaluation refers to “[…] the expression of the speaker’s or writer’s attitude or stance towards, viewpoint on, or feelings about the entities or propositions [statements] that he or she is talking about. That attitude may relate to certainty or obligation or desirability or any of a number of other sets of values” (Hunston – Thompson 2000, p. 5). Evaluation expresses the speaker/writer’s opinions, thus reflecting their value systems and those of their community; it serves to construct relationships between speakers and readers; and it helps to organise texts (Hunston – Thompson 2000).

The appraisal system developed by Martin and White (2005) further contributed to clarify the function of the language of evaluation in the LEAQ corpus. The features of the commentator voice (judgement, affect, appreciation) used to either condemn or praise, and their associated values of positivity/negativity were particularly useful to understand the evaluative stance of *The Times* on the matters discussed in the letters. These were put in relation with further studies on how corpora are used to conduct studies on evaluation and evaluative phraseology in a variety of text types (Hunston 2011; Gozdz-Roszkowski – Hunston 2017). Phraseology, as pointed out by Hunston (2011, p. 5) “describes the general tendency of words, and group of words, to occur more frequently in some environments than in others”. Therefore, studying the co-text, i.e., the environment, of evaluative lexical items and their collocates and clusters helped to better understand the textual strategies of the LTE making up the LEAQ corpus.

When applied to news discourse, the study of the language of evaluation can be used to understand the evaluative stance of the news institution, how it reflects its news values, i.e. what makes something newsworthy, its relationship between readers and news writers, and its way of organising news stories (Bednarek 2010). Using a corpus approach to study the evaluative language in the news, parameters of evaluative language have been identified that contribute to newsworthiness (Bednarek 2006; Bednarek 2010) and eventually conflated in the Discursive News Value

Analysis (DNVA), an approach developed by Bednarek and Caple (2017; 2019) to understand how newsworthiness is constructed through different semiotic sources.

A corpus-driven (Tognini-Bonelli 2001; Sinclair 1996, 2004) quantitative and qualitative approach allowed me to identify the most recurrent evaluative adjectives out of the general word list obtained with WordSmith Tools 8.0 (Scott 2020) of the LEAQ corpus. Evaluative adjectives are used to express the position writers take towards their content, and they serve as an explicit or implicit signal of their stance. Therefore, they could be regarded as linguistic items that are frequently used to influence the perception of readers on a certain news item. The corpus-driven analysis provided quantitatively relevant evaluative adjectives; the analysis of their most frequent concordances and collocates followed, without any preconceived concepts orienting the choice of the items to be analysed apart from their frequency in the corpus. A corpus-driven approach is particularly relevant in this research, because it allows data to emerge directly from the analysis of the corpus.

5. De-constructing newsworthiness through the analysis of the language of evaluation

The Letters to Editor on the Armenian Question (LEAQ) corpus was collected from the digital online archive of *The Times and The Sunday Times*. Hosting the complete collection of the articles published between 1785 and 1985 matches the standard of completeness in corpus building (Hunston 2002). The letters were selected using two significant search words, *Armenia* and *Armenian*, the latter including also letters where the term *Armenians* occurs. This resulted in collecting all the letters to the editor where the Armenian question was mentioned over a span of twelve years, from 1914 to 1926. This span of time was selected to also attempt a reconstruction of the context immediately before the onset of the genocide and after, and to see how and if any linguistic signals could be detected that could somehow anticipate the events of 1915.

The LEAQ corpus amounts to 186 letters for a total of around 120,000 tokens. The letters were downloaded in both PDF and OCR formats; the OCR files were edited and compared with corresponding PDF files to ensure correctness, renamed with their date and page of publication, and saved as UTF-8 TXT files. Digitised files were then processed using WordSmith Tools v.8.0 (Scott 2020) to obtain a wordlist out of which the most recurrent evaluative adjectives were isolated; due to the limited number of texts featured in the corpus, the selection could be done manually. Table 1 exemplifies the most frequently occurring evaluative adjectives:

Table 1. Most frequently occurring evaluative adjectives in the LEAQ corpus

N		Word	Freq.	N		Word	Freq.
1	184	POSSIBLE	69	11	457	STRONG	30
2	245	GOOD	55	12	471	SUPREME	29
3	253	RECENT	53	13	492	INDEPENDENT	28
4	264	CERTAIN	51	14	519	SIMILAR	26
5	272	OLD	50	15	525	OFFICIAL	26
6	284	OBEDIENT	47	16	529	COMPLETE	26
7	288	KNOWN	47	17	550	HIGH	25
8	294	NECESSARY	46	18	560	TERRIBLE	24
9	302	LONG	45	19	578	COMMON	24
10	417	IMPORTANT	33	20	621	IMPOSSIBLE	22

The first recurrent evaluative adjectives (*possible*, *good*, *recent*, *certain*, *old*) could be ascribed to different parameters taken from the classification by Bednarek (2010) (possibility, positivity, recency or timeliness, unambiguity, and again recency or timeliness), which expands and further defines Hunston and Thompson (2000) and Martin and White (2005). However, other parameters could be attributed to the results from the key word list, namely necessity (*necessary*, *essential*), emotivity (*terrible*, *unfortunate*, *disastrous*), importance (*important*), expectedness (*certain*, *known*, *clear*, *expected*), as well as comparators (*different*), following the work of Hunston and Thompson (2000), or unexpectedness (*different*), following again the most recent work by Bednarek and Caple (2019). Often, however, more parameters are applicable to the same adjective, depending on the various evaluative meanings associated to the adjective itself and depending on its context of use.

Parameters from different studies by Bednarek (2006, 2010) were used, as her recent works with Caple (Bednarek – Caple 2017, 2019) draws and selects from her more extensive set of parameters; also, some recurrent adjectives, such as the most recurrent adjective *possible* was difficult to fit into her latest selection of parameters per se (consonance, eliteness, impact, negativity, personalisation, proximity, superlativeness, timeliness, unexpectedness). These evaluative parameters are used to analyse media discourse in the new and are here applied instead to analyse LTE.

For the limited scope of this article, the analysis is focused on the most recurrent evaluative adjective *possible* and on its different evaluative meanings in predicative and attributive grammatical structures (Biber et al. 2007), following Samson (2006), to study its occurrences in the ideally "unmediated authorially sourced judgement" (Martin – White 2005) of the LTE of the LEAQ corpus.

The study of *possible* allows one to understand how the newsworthiness and relevance (Wahl-Jorgensen 2002) of the topic was construed in the LTE using the news value parameter of superlativeness (Bednarek 2010; Bednarek – Caple 2017, 2019), and the concept of evaluative connotational meaning as outlined in Morley and Partington (2009) and relying on semantic prosody (Sinclair 2003), in relation also to lexico-grammatical collocates pertaining to the grammar of modality (Halliday – Matthiessen 2014). Further research activity is already planned to build on the results of the analysis presented in this article, and to examine other more and less frequently occurring evaluative adjectives in the LEAQ corpus in order to contribute to the study of the local grammar of evaluation and of the linguistic and textual features of the letters to editor.

5.1. Possible – attributive use

As previously anticipated, among the evaluative parameters singled out by Bednarek and Caple (2017, 2019), *possible* is not clearly ascribed to one of the news values conferring newsworthiness. However, its leading position in the LEAQ corpus needs a more in-depth analysis to understand the reasons behind its frequency. *Possible per se* might be associated to the parameter of possibility and of reliability as formulated by Bednarek (2010) in her methodological framework of evaluation in the news, following Hunston and Thompson (2000) in connection with the evaluative parameter of certainty, and to hedging and its related aspects of modality (Martin – White 2005; Hunston 2011; Halliday – Matthiessen 2014). Among the four rules for selection of the content of LTE, namely relevance, brevity, entertainment, and authority (Wahl-Jorgensen 2002), the rule of relevance to the events and the rule of authority are those along which the LEAQ corpus seems to be organised. In view of its small size, an overall individual reading of the texts was indeed possible. Also, thanks to WordSmith Tools 8.0 (Scott 2020), it is possible to add a diachronic perspective to the analysis, to verify if changes in the evaluative connotational meaning of *possible* occurred in the span of time under consideration, in view of the evolving events surrounding the Armenian question.

Concordances for *possible* in the LEAQ corpus can be automatically listed in ascending chronological order, following file naming with the day and page of publication of each letter to the editor; therefore, Table 2 shows the first concordances of *possible* appearing in the corpus.

Table 2. Chronologically first occurrences of *possible* in the LEAQ corpus

provided for. These are being cared for as far as **possible** for the moment by the Russian Armenian inhabitant
ocal committees, are rendering all the assistance **possible,** but they have no funds left, all the money subscr
thorities are separating the fugitives as much as **possible,** as it is feared there may be an outbreak of disea
f Easterns, I should like to state as strongly as **possible** that the inhabitants of the Ottoman dominions, be
nt in the Ottoman dominions. It is, however, just **possible** that their repetition in a letter to The Times ma
in the conduct of Balkan affairs. It is not only **possible,** but highly probable, that mistakes may have been
plete change of Ministers. I dare say it would be **possible** for a partisan politician, or even for one not an
ve still to learn that such redress as may yet be **possible** has been made for that act of murder. Americans a

These first occurrences all appear in 1915. More specifically, the first three occurrences appear in the same letter published on 12 January 1915, some months prior to the actual start of the massacres in April in that same year. It is worth remembering that the selection of the span of time preceding the actual massacres was intended to detect some potential signs that the Armenian massacres were possibly anticipated by other events reaching the news. This appears to be the case in this first letter, titled “The Armenian Red Cross”, where the evaluative adjective *possible* first occurs in the sentence in Example 1:

(1) “These are being cared for as far as possible for the moment by the Russian Armenian inhabitants, who are themselves very poor owing to floods having spoilt their last crops”.

This first occurrence shows an attributive structure whereby *possible* is pre-modified by a comparative adverbial structure, that is repeated in other subsequent occurrences. The anaphoric reference of the deictic subject pronoun *these* is to be found in the short preceding sentence, "There are now 12,000 Armenian refugees at Sarikamysch alone to be provided for." Who were those refugees? Why were they refugees? What were they trying to escape? Further information is added on the conditions of the refugees after Example 1, explaining why they are being cared: "Hundreds of old men, women, and children have tramped through the snow without shoes or stockings, these articles having been seized by Turkish soldiers, who had been billeted in their houses".

This first occurrence of *possible* is part of a comparative adjectival structure whereby a sense of limitation, or a sense of reaching a limit of achievability is expressed. This same evaluative sense is also conveyed by the other two occurrences of *possible* inside this first letter, that is to say *all the assistance possible* and *as much as possible*. Example 2 and example 3 provide the context where they occurred:

(2) "The Catholices (head of the Armenian Church) and his clergy, with local committees, are rendering all the assistance possible, but they have no funds left, all the money subscribed by Armenians having to go to the upkeep of the volunteers".

(3) "The Russian authorities are separating the fugitives as much as possible, as it is feared there may be an outbreak of disease, owing to their famished and impoverished conditions".

In Example 1 and Example 3 above, both attributive adverbial phrases are post-modifying a verb phrase ("[The refugees] are being cared of" and "are separating [the fugitives]"), in which the action expressed by the verb seems to reach its limit of achievability with both evaluative adjective phrases, implying also, to some extent, a limitation of responsibility, due, in this case, to the lack of funds. Therefore, if efforts have reached their limit of feasibility, then it might be argued that whoever is responsible for making those efforts is somehow discharged of the responsibility towards the need to make more efforts, seemingly having fulfilled their responsibilities at the same time. The same paradoxical connotation is expressed in Example 2 through the structure *quantifier + article + noun + possible* (as much as possible). The connotation assigned to these attributive structures relying on adverbial comparisons and quantifiers is used to convey the evaluative meaning of limitation of feasibility, which, together with a limitation to responsibility seems, however, also a call for receiving help.

The same connotation is also expressed in one of the most frequent collocates of *possible*, the quantifier *every*, located in its immediate L1 position. Table 3 below reports the results in context of the collocational pattern *every* + *possible*:

Table 3. Attributive collocational pattern every + *possible*

ed to our country, I will continue to help in **every possible** way as I have done in the Senate in the last two
he mandate, while those same Powers imposed **every possible** restriction on its own action, going so far as to
ould allow such licence to an avowed enemy. **Every possible** means should be employed to combat the inference
le for massacres of defenceless Christians. **Every possible** means should be taken to indicate to these bloodt
d to be able to assure our subscribers that **every possible** precaution is taken that our gifts shall reach th
al that we should work towards that goal by every **possible** means. As for present-day Russia (continues the a

The phrase *every* + *possible* further collocates with nouns (along the structure *quantifier* + *adjective* + *noun*), of which the most frequent is *every* + *possible* + *means*; the same cluster *every* + *possible* collocates in turn with *way*, *precaution* and *restriction*. Example 4 shows one example in its context:

(4) Every possible means should be taken to indicate to these bloodthirsty outlaws of the centuries that Christian civilised men will not shake hands with them, or have any sort of intercourse with them.

The use of the quantifier in Example 4 and in Example 2, but also the adverbial comparisons in Example 1 and 3 are clearly related to the news value parameter of superlativeness, according to which "the event is constructed as being of high intensity or large scope/scale" (Bednarek – Caple 2019, p. 93), the extent of which can be "established through the linguistic resources of intensification and quantification" (Bednarek – Caple 2019, p. 93). Therefore, intensifiers such as *as much as*, *as far as*, *all the*, *every* contribute to the newsworthiness of *possible* in its attributive construction by recurring to an intensification of the event to which these attributive occurrences of *possible* are related.

5.2. Possible – predicative use

The evaluative meaning of *possible* conveyed through its predicative use in the other first six occurrences from 1915 shown in Table 2 above, and occurring in four different letters, relies instead on the grammar of modality and on gradability, and is differently connoted. Example 5 shows the fifth occurrence in its context, from a letter by Lord Cromer titled *Germany and the East. Lord Cromer's Warning*. Actually, within the context of the letter, this is the first occurrence of *possible*, despite the fact that in the results provided by WordSmith Tool v.8.0 (Scott 2020) this occurrence is listed as fifth occurrence, and not as fourth, as it should be.

(5) It is, however, just possible that their repetition in a letter to The Times may arrest the attention of some who are interested in Eastern affairs and who are fortunate enough to be living for the time being in countries which admit of the circulation of news and of opinions. (The Times, 30 July 1915, p. 7)

Among the letters of the LEAQ corpus, this one is not included in the collection by Peltekian (2013), and it reports a reply by Lord Crewe in the House of Lords which may "arrest the attention" of the readers, appealing to their interest and to their solidarity, as well as to their deeper understanding of political dynamics underlying the responsibility of the German army, that did not interfere with the "wholesale massacre and deportation" carried out in Armenia. In this case, *possible* is used within a predicative construction to suggest the chance of this piece of news, otherwise restricted to the House of Lords, to be spread and to be trusted, particularly because it is authored by an authoritative voice. Moreover, this structure corresponds to the first pattern of the grammar of evaluation mentioned by Hunston and Sinclair (2000) to recognise evaluative adjectives (*it+ link verb + adjective group + clause*), which is also the same structure of the occurrences 6 to 8 of Table 2.

Example 6 shows the use of the evaluative adjective in the same letter, integrating the grammar of modality of a predicative structure with an attributive comparative adverbial structure (*as strongly as*), to reinforce the strength of the authorial voice:

(6) As one who has passed the best years of his life in the East and takes the deepest interest in the moral and material welfare of Easterns, I should like to state as strongly as possible that the inhabitants of

> the Ottoman dominions, be they Moslem or Christian, have nothing whatever to hope from the establishment of German predominance in their midst.

With the evaluative adjective *possible* being pre-modified by a comparative adverbial structure, qualifying the verb *state*, the intention behind the use of such a structure is to claim a strong position, a challenging political position that was relevant to the target reading audience of the letters in the historical context of World War I. The Armenian genocide, here, is used to reinforce accusations towards the German enemy, by adding a complementary perspective to the context of World War II.

The analysis of the left- and right-collocates of *possible* adds further insight into how it was used to express evaluative meanings inside the LEAQ corpus. Most frequent collocates in R1 position are *that*, *for*, while *every* is the most recurrent collocate in L1 position, as discussed before. As shown in Table 3 below, the collocational pattern with *that* highlights the predicative use of *possible*, both in the positive structure *it + link verb + adverb + possible + that*, and in the cluster based on the interrogative structure *link verb + it + possible + that*. The positive structure, instead, shows that the affirmative strength of *possible* is often graded through an adverbial pre-modification, such as with *quite* or *just*, making it similar to its use with modal verb phrases (*may*, *would*) or with future verb phrases (*will*). Table 4 shows the collocational pattern with *for* in R1 position:

Table 4. Collocational pattern *possible + for*

large areas of Europe and Asia Minor. Would it be **possible for** you to write a letter, either to myself or ot
including 150,000 refugees from Asia Minor. Is it **possible for** public opinion in Great Britain indifferently
provided for. These are being cared for as far as **possible for** the moment by the Russian Armenian inhabitant
." If America rejects them " it will no longer be **possible for** America to exercise effective influence in th
ned to irresponsible articles and speeches it was **possible for** moderate Mahomedans in India to argue that th
iversities are also crammed to overflowing. Is it **possible for** the Government, who represent the ratepayer a

plete change of Ministers. I dare say it would be **possible for** a partisan politician, or even for one not an
passed and nothing has been done. (2) To make it **possible for** the Armenians of Van, &c., who are now crowde
ower in the hands of this country it would not be **possible for** Constantinople, lying under the guns of the A

In this case, *possible for* collocates with place names used as personifications (Constantinople, America), a national group (Armenians), a religious group (moderate Mahomedans), socio-political lexical items (a partisan politician, the Government, public opinion), and to form a prepositional phrase of time (for the moment). In most cases then, *possible for* is constructed as *adjective* + *for* (+ *article*) + *noun*, with the noun introducing another term into the discourse, who is potentially in charge of performing a certain action, as shown in Example 6, allowing a future course of action (Hunston – Thompson 2000) and hinting at an ideational metafunction (Halliday – Matthiessen 2014):

(6) Is it possible for public opinion in Britain indifferently to envisage the further destruction of so many homes and live, and fortunes, amounting to many hundreds of millions?

This example, from a letter by E.K. Venizelos, i.e. Eleftherios Kyriakou Venizelos, a prominent leader of the Greek national liberation movement, also mirrors the most recurrent cluster, *link verb* + *it* + *possible* + *that*, with the that-clause replaced by a prepositional phrase (for + noun) delaying the to-infinitive clause, and emphasising the components of the clause (the prepositional phrase *in Britain* and the evaluative adverb phrase *indifferently*). Analysing also other occurrences of *possible* + *for* + *noun*, and particularly its left and right collocates, it is evident that the same interrogative structure emerges whenever a call to action is claimed.

The predicative structure of *possible* in the LEAQ corpus, therefore, seems to occur to question a certain course of action, and to instil in the audience an element of doubt, which is mostly what the grammar of modality achieves. Indeed, by recalling potentiality (Halliday – Matthiessen 2014), the adjective *possible* adds explicit and high polarised subjective evaluation that convey the writer's position, despite the scarcity of evaluative adjectives retrieved in the LEAQ corpus.

As seen in the examples, the use of the evaluative adjective *possible* in the LEAQ corpus does not comply with the news value parameter of possibility but tends to be associated more with the news value parameter of superlativeness in its attributive form, and to the grammar of modality in its predicative form, with an overlapping of the two when reinforcement of the evaluative meaning intended to convey is needed. The attributive structure, with the use of comparisons and quantifiers, conveys a sense of limitation, as if the limit of feasibility concerning the noun of the attributive structure has been somehow reached, and nothing else can be done – apart from veiled call for help aimed at the readership, or through the readership, of the letters, particularly when *possible* left-collocates with the quantifier *every*. Instead, the evaluative meaning associated to the predicative form of *possible* acquires a moral connotation, conveying the writer's opinion on what should have morally been done, particularly after the onset of the genocide.

It is indeed not only the alternation of attributive and predicative evaluative meanings, but also the chronological distribution of these local structures which adds to the local grammar of evaluation of the corpus. These types of evaluation seem to occur in two different moments in relation to the events of the Armenian Question. The attributive meaning of *possible*, with its intrinsic value of limitation, seems to occur at the same time of the massacres, or in the immediate aftermath. The predicative meaning of *possible*, instead, with the moral accusation implied by its collocates and by the contexts where it is featured, seems to occur more frequently in later years, as strong criticism of what has been done, or of what has not been done, until then. However, further extensive research should be done in order to verify more accurately these trends.

An ambivalent use of the evaluative adjective *possible* therefore appears to be in use, with an evolution from an evaluative meaning of limitation in its attributive use, to an evaluative meaning critical of the current situation and envisaging future courses of action in its predicative use. The discursive news value of the evaluative adjective *possible* not only relied on two opposing meanings, but, in view of its frequency, is a dominant value of the news discourse inside LEAQ, insofar as it creates a structural evaluative ambivalence. Therefore, the attributive use of *possible* is then related to making conclusive statements, whereby, if everything possible has been done, the meaning attached to this evaluative use of possible leads to a general discard of responsibility. Blame, or at least a moral connotation is expressed by the predicative use of *possible*, according to a dialectic of disclaiming *vs*. claiming, discharging responsibility *vs.* charging with responsibility, that alternates throughout the corpus.

6. Concluding remarks

On the basis of the analysis performed so far, the LEAQ corpus shows features that possibly contribute to understand the linguistic reception and subsequent acknowledgment of the Armenian genocide. LTE are usually based upon a reaction to certain news items, but, in turn, they contribute to generate a reaction in the audience, according to how their newsworthiness is linguistically constructed through the use of evaluative language. The two different evaluative meanings emerging from the attributive and the predicative collocational structures of *possible* are organised along a polarised continuum between limitation of achievement and blame for lack of achievement, expressed through the attributive collocates (quantifiers, comparison) and through the predicative collocates (affirmative and interrogative forms, that-clause, prepositional phrase *for+noun*), which, sometimes, are also blended inside the same sentence.

Limiting the scope of the analysis to one single evaluative adjective and to its collocational patterns provided a significant example of the evaluative lexico-grammatical structures that in the LEAQ corpus contribute to the linguistic features of LTE that construct newsworthiness. However, the textual construction of newsworthiness in the LTE of the LEAQ corpus needs to be examined further to better understand which of its linguistic features were most used to influence the perception of the reading public. The polarised continuum between limitation of achievement and blame for lack of achievement identified with the analysis of possible seems indeed to suggest an underlying collective perception of the events that might emerge when extending the analysis to other recurrent evaluative adjectives. Isolating further lexico-grammatical features that contribute to the construction of newsworthiness in the corpus would also help to better understand whether some linguistic strategies adopted by the international press, somehow, might have contributed to the oblivion of the Armenian genocide. Ultimately, the LEAQ corpus represents not only a sample of the public debate on the events surrounding the Armenian genocide, but also an example of the language of LTE in use around the first decades of the 20th century in a British broadsheet newspaper.

REFERENCES

Sources

The Times and The Sunday Times Online Archive
https://www.thetimes.co.uk/archive/, accessed December 2021

Special studies

Alayrian, A.
2018 *Consequences of Denial. The Armenian Genocide*. London and New York: Routledge.

Astourian, S.
1990 "The Armenian Genocide: An Interpretation", *The History Teacher* 23 (2), 111-160.

Aybak, T.
2016 "Geopolitics of denial: Turkish state's 'Armenian problem', *Journal of Balkan and Near Eastern Studies* 18 (2), 125-144.

Biber, D. et al.
2007 *Longman grammar of spoken and written English.* Edinburgh: Pearson Longman.

Bednarek, M.
2006 *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*. New York / London: Continuum.

Bednarek M.
2010 "Evaluation in the news. A methodological framework for analysing evaluative language in journalism", *Australian Journal of Communication* 37 (2), 15-50.

Bednarek M. – H. Caple
2017 *The Discourse of News Values: How News Organizations Create Newsworthiness*. Oxford: Oxford University Press.

Bednarek M. – H. Caple
2019 *News discourse*. London / New York: Bloomsbury Academic.

Brownlees N. et al. (eds.)
2010 *The Language of Public and Private Communication in a Historical Perspective*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Cavanagh A.
2019 "Letters to the Editor as a Tool of Citizenship". In: Cavanagh A. – J. Steel (eds.) *Letters to the Editor. Comparative and Historical Perspectives*. London: Palgrave Macmillan, 89-108.

Chabot J. et al.
2016 *Mass Media and the Genocide of the Armenians*. London: Palgrave Macmillan.

Chovanec, J.

2012 "From adverts to letters to the editor. External voicing in early sports match announcement". In: Palander-Collin, M. et al. (eds.) *Diachronic Developments in English News Discourse*. Amsterdam: John Benjamins, 175-197.

Elayyadi H.

2017 "Reconciliation Process. How has the Turkish state's official discourse of the Armenian Genocide evolved during the Erdogan era?". In *Armenian Journal of Political Science* 2 (7), 77-94.

Gozdz-Roszkowski S. – S. Hunston

2017 "Corpora and beyond – investigating evaluation in discourse: introduction to the special issue on corpus approaches to evaluation", *Corpora* 11 (2), 131-141.

Halliday M.A.K. – C. Matthiessen (eds.)

2014 *Halliday's Introduction to Functional Grammar. Fourth Edition*. London / New York: Routledge.

Hobbs A.

2019 "Readers' Letters to Victorian Local Newspapers as Journalistic Genre". In: Cavanagh A. – J. Steel (eds.) *Letters to the Editor. Comparative and Historical Perspectives*. London: Palgrave Macmillan, 129-146.

Hunston S.

2002 *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston S.

2011 *Corpus Approaches in Evaluation. Phraseology and Evaluative Language*. New York: Routledge.

Hunston S. – J. Sinclair

2000 "A Local Grammar of Evaluation". In: Hunston S. and Thompson G. (eds.), *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press, 75-101

Hunston S. – G. Thompson (eds.)

2000 *Evaluation in Text: Authorial Stance and the Construction of Discourse*, Oxford University Press, Oxford.

Mamali C. et al.

2019 "Conflicting representations on Armenian genocide: exploring the relational future through self-inquiring technique", *Filozofia Publiczna i Edukacja Demokratyczna* 2 (8), 168-250.

Martin J.R. – P.R.R.White

2005 *The Language of Evaluation. Appraisal in English*. London: Palgrave Macmillan.

Mayersen D.

2016 *On the Path to Genocide: Armenia and Rwanda Reexamined*. New York & Oxford: Berghahn.

Morley J. – A. Partington

2009 "A few Frequently Asked Questions about semantic – or evaluative – prosody", *International Journal of Corpus Linguistics* 14 (2), 139-158.

Peltekian K.M.

2013 *The Times of Armenian Genocide. Reports in the British Press. Volume 1: 1914-1919*, *Volume 2: 1920-1923*. Beirut: Four Roads.

Pounds G.

2005 "Writers argumentative Attitude: A contrastive analysis of 'Letters to the Editor in English and Italian", *Pragmatics* 15 (1), 49-88.

Pounds G.

2006 "Democratic participation and Letters to the Editor in Britain and Italy", *Discourse & Society* 17 (1), 29-63.

Richardson J.E. – B. Franklin

2004 "Letters of Intent: Election Campaigning and Orchestrated Public Debate in Local Newspapers' Letters to the Editor", *Political Communication* 21, 459-478.

Romova Z. – J. Hetet

2012 "Letters to the editor: Results of Corpus Analysis", *New Zealand Studies in Applied Linguistics* 18 (2), 45-63.

Samson C.

2006 "'… is different from…'. A corpus-based study of evaluative adjectives in economics discourse", *IEEE TRANSACTIONS ON PROFESSIONAL COMMUNICATION* 49 (3), 236-245.

Sinclair J.

1996 "The Search for Units of Meaning", *TEXTUS* IX (1), 75-106.

Sinclair J.

2003 *Reading Concordances: An Introduction*. London: Longman.

Sinclair J.

2004 *Trust the Text. Language, Corpus, and Discourse*. London: Routledge.

Scott M.

2020 *WordSmith Tools version 8*, Lexical Analysis SoftwareStroud.

Tognini-Bonelli E.

2001 *Corpus Linguistics at Work*. Amsterdam/Philadelphia: Benjamins.

Wahl-Jorgensen K.

2002 "Understanding the Conditions for Public Discourse: four rules for selecting letters to the editor", *Journalism Studies* 3 (1), 69-81.

Address: Isabella Martini, Università degli Studi di Firenze, Dipartimento di Formazione, Lingue, Intercultura, Letterature e Psicologia, Stanza 211, Via Santa Reparata, 93/95, 50129 Firenze, Italy.
ORCID code: https://orcid.org/0000-0002-6086-3313

Comparison of key statistical instruments used in lexicon-based tools for sentiment analysis in the English language

Łukasz Stolarski

Jan Kochanowski University of Kielce

ABSTRACT

This study investigates "key statistical instruments", such as the mean or the sum, used in obtaining numeric polarity scores in lexicon-based tools for sentiment analysis. First, a large number of texts rated for sentiment intensity by independent human judges was collected. Next, 15 different sentiment lexicons were used to generate sets of numeric values for each of the texts. Then, the key statistical instruments were calculated on the basis of these results and compared with the corresponding human scoring using tests for association between paired samples. The results of these tests were further examined with the use of ANOVA and Tukey HSD post-hoc analysis. The broad conclusion drawn from the analysis is that the mean, all other things being equal, is the most reliable key statistical instrument for obtaining numeric polarity scores that are similar to scores provided by human assessors. These results may be of particular importance for both developers of lexicon-based programs performing sentiment analysis and users of such software packages.

Keywords: sentiment analysis, opinion mining, sentiment mining, opinion extraction.

1. Background

Sentiment analysis is defined as "the polarity of an opinion item which either can be positive, neutral or negative" (Borth et al. 2013: 223) or a procedure which involves "determining the evaluative nature of a piece of text" (Kiritchenko et al. 2014: 723). More specifically, the term is typically used in reference to "an active area of study in the field of natural language processing that analyses people's opinions, sentiments, evaluations, attitudes, and emotions via the computational treatment of subjectivity in

DOI:10.25951/4827

text" (Hutto – Gilbert 2014: 217) or, in short, to "the computational treatment of opinion, sentiment, and subjectivity in text" (Pang – Lee 2008: 10). The subject has been explored in a large number of publications, and several reviews of literature on sentiment analysis are available (e.g. Liu 2012; Liu – Zhang 2012; Pang – Lee 2008).

Tools for performing sentence-level sentiment analysis are frequently divided into two major categories. The first one, which may be referred to as "the machine learning approach" (Ribeiro et al. 2016; Taboada et al. 2011), involves labelled training data which are used for building a classifier. Such tools are used for conducting sentiment analysis on particular types of texts, as they usually perform very well in the domain that they were trained on. Nevertheless, their performance may drop considerably in other domains (Aue – Gamon 2005) and they do not cope well with effects of linguistic context such as negation or intensification (Taboada et al. 2011: 269). "The lexicon-based approach", on the other hand, makes use of a list of words, or "a sentiment lexicon", in which each word or phrase is assigned a sentiment value. Some such lexicons involve categorical classification. An example of this is the NRC Emotion Lexicon (Mohammad – Turney 2010, 2013), which contains, among other categories, the binary distinction between "positive" and "negative". Other lexicons offer continuous polarity scores, as is the case for all the lexicons described in Section 3.2. Sentiment lexicons also differ in the way they are obtained. Some are created manually, usually involving a group of participants whose task is to label selected words in terms of sentiment polarity or value. Such tasks tend to be costly, time-consuming and labour-intensive; hence, the resulting lexicons are relatively small. Typically, they contain a few thousand words. However, they tend to be less domain specific and the tools that utilize them are usually more consistent across domains (Taboada et al. 2011). Examples of lexicons which involve human annotation are the aforementioned NRC Emotion Lexicon, Sentiment Composition Lexicon for Negators, Modals, and Degree Adverbs (SCL-NMA) described in Section 3.2.6, as well as, at least partially, the lexicons used in SentiStrength and Vader projects (see Sections 3.2.2 and 3.2.5, respectively). A large number of sentiment lexicons are, nevertheless, created automatically using seed words. They tend to contain a greater number of unigrams and sometimes longer expressions (bigrams and trigrams), but their performance may be less consistent across domains. Most of the lexicons described in Section 3.2.6 were created in this way.

Several benchmark comparisons of sentiment analysis tools have recently been published (e.g. Abbasi et al. 2014; Diniz et al. 2016; Gonçalves et al. 2013; Ribeiro et al. 2016). They demonstrate that, on average, some software packages perform better than others; however, there is no clear winner for all possible testing sets. The performance of individual sentiment analysis tools varies depending on the domain which is being investigated.

It should be noted that studies on sentiment analysis frequently discuss potential problems which may affect results. The most pressing issues include negation and intensification. These two aspects have received much attention and numerous solutions have been suggested. For example, it was initially proposed that negation could be dealt with by reversing the polarity of a lexical item (Choi – Cardie 2008; Kennedy – Inkpen 2006). This approach, however, has been shown to be fundamentally flawed (Kennedy – Inkpen 2006; Kiritchenko et al. 2014; Taboada et al. 2011); thus, alternative solutions, such as shifting the polarity by a fixed amount, have been used (Taboada et al. 2011). A useful taxonomy of problems affecting sentiment analysis is offered in Abbasi et al. (2014). It demonstrates that even though the performance of some tools may be promising, there is still much room for improvement, and further research is necessary.

2. Aims

This paper focuses on lexicon-based tools that perform sentiment analysis on phrases, sentences and longer utterances and give continuous polarity scores. When using such tools, it becomes clear that they consist of two largely independent components. The first is a sentiment lexicon, or a group of such lexicons. The second is a sentiment analysis algorithm which calculates the final score on the basis of several (modified) digits representing sentiment values of individual words or phrases. These values may simply be added, but other solutions are also possible, e.g. calculating the mean, median or obtaining the highest absolute value.

The major aim of this project is to compare the effectiveness of such key statistical instruments in calculating final sentiment scores (for the description of the exact methods tested see Section 3.1). They are necessary at the final levels of sentiment analysis performed by tools within the lexicon-based approach. Consequently, investigating the efficacy of such statistics, other things being equal, may help in improving the overall performance of sentiment analysis tools. Additionally, the results of this study may also be useful for the end users of such software packages. In some cases, the user may choose between various ways in which the final sentiment score is calculated (e.g. SentiStrength).

It must be stressed that this study is not a benchmark comparison of any software packages. Rather than testing actual sentiment analysis tools, the current analysis focuses on the efficacy of key statistical instruments applied to "bare" sentiment lexicons. The resulting correlation with human scores is expected to be lower than the corresponding correlation obtained using complete software packages for sentiment analysis. Such packages may involve various additional strategies to deal with the problems mentioned in Section 1. Nevertheless, testing key statistical instruments on "bare" sentiment lexicons is fundamental to lexicon-based sentiment analysis and, as suggested in the previous paragraph, may be essential in improving the performance of actual software packages.

3. Methods

In order to accomplish the major aim outlined in Section 2, five key statistical instruments were defined (see Section 3.1). Next, a group of sentiment lexicons with continuous polarity scores was selected (see Section 3.2). After that, a representative number of validation texts with numerical scores for the positive-negative dichotomy provided by human respondents was obtained (see Section 3.3). Then, the key statistical instruments for each validation text were calculated on the basis of each sentiment lexicon. This task involved some text preprocessing. Each case required a slightly different approach and the details on the preprocessing are provided in the description of each lexicon. Finally, statistical tests were performed on the data obtained (see Sections 3.4 and 4).

3.1 The key statistical instruments

The five key statistical instruments chosen for comparison are summarised below.

- MEAN1 – the mean obtained on the basis of all scores, excluding the value of 0.0 added to lexical items not recognized in a given lexicon or stop words removed from the analysis. In the example presented in Table 1, the mean would be calculated as follows:
  $(0.9765 + 0.7181 - 0.3638 - 1.4753) / 4 = -0.0361$.
- MEAN2 – the mean obtained on the basis of all scores, including the value of 0.0 added to lexical items not recognized in a given lexicon or

stop words removed from the analysis. The result for the example in Table 1 would be determined in the following way:
(0.9765 + 0.7181 – 0.3638 – 1.4753) / 10 = –0.0144.

- MEDIAN – the median obtained on the basis of all scores, excluding the value of 0.0 added to lexical items not recognized in a given lexicon or stop words removed from the analysis. For the example in Table 1, MEDIAN = 0.1775.
- LAV (largest absolute value) – the largest value in all the scores, regardless of the polarity. For the example in Table 1, LAV = –1.4753.
- SUM – the sum obtained from all the scores. For the example in Table 1, SUM = –0.1445.

MEAN is a measurement which could easily be applied in obtaining final scores for sentiment intensity in lexicon-based tools. It is offered as one of several options in the GUI distribution of SentiStrength (see Section 3.2.2). This statistical instrument will be investigated in the two versions described above to see if the inclusion/exclusion of elements with no sentiment scoring affects the results. MEDIAN, to the best of the author's knowledge, has not been used in sentiment analysis software packages, but is appropriate in this study. It has characteristics similar to those of MEAN, as its purpose is to summarise datasets, but it is less affected by outliers. By contrast, LAV represents only the most extreme value in a dataset. This statistical instrument is offered as one of the options in the GUI distribution of SentiStrength. Finally, SUM is probably the most obvious solution applied in the calculation of final scores. It is used, for instance, in Vader (see Section 3.2.5) and Afinn (Nielsen 2011).

In addition to the above statistics, other possible calculations were also considered. For instance, MEDIAN could also have been calculated on the basis of all scores, including the zeroes assigned to items not recognized in a lexicon or stop words. However, the results yielded 0.0 in too many cases; thus, the method was considered significantly less reliable than the other five. Additionally, "mode" was also excluded from the analysis because it is not appropriate for continuous data.

Table 1. Example results for a text containing both positive and negative lexical items

	he	is	friendly	and	funny	but	also	naive	and	irresponsible
scores	0.0	0.0	0.9765	0.0	0.7181	0.0	0.0	–0.364	0.0	–1.4753

3.2 Sentiment lexicons

Fifteen different sentiment lexicons were used in this study. All of them involve numerical scoring. Five are associated with independent projects (SenticNet, SentiStrength, SentiWordNet, UMass Amherst Linguistics Sentiment Corpora, Vader), nine belong to the set of sentiment lexicons created by the National Research Council Canada (NRC) and the last one was created on the basis of these nine lexicons (see the last paragraph on "NRC Combined" in Section 3.2.6). All the lexicons are described in Sections 3.2.1 to 3.2.6. In each case, a general summary is presented and the way a given lexicon was used in the present study is summarized.

3.2.1 SenticNet

SenticNet (Cambria et al. 2016) is a project conceived at the MIT Media Laboratory in 2009. Its development involves collaboration between the Media Lab, the University of Stirling, and Sitekit Solutions Ltd. It is accessible by an API available online, but the exact tool used in the present study is the Python package "senticnet" (ver. 1.0.1).

The package is not just a sentiment lexicon. It offers several useful options. They are available mostly for individual words, but some phrases may also be queried. Among other things, one may obtain "moodtags", such as "#joy" or "#admiration", or the so-called "sentics", which are values for qualities such as "sensitivity", "attention", "aptitude" and "pleasantness". Most relevant to the present study, however, are the attributes "polarity value" and "polarity intensity". The former is a binary sentiment value for a given word (positive or negative), and the latter refers to a similar result on a gradable scale of –1 (extremely negative) to +1 (extremely positive). "Polarity intensity" is, therefore, the feature which has been used for the current purposes.

The application of SenticNet into the analysis described in Section 4 is fairly straightforward (see Figure 1). Each text from the validation materials described in Section 3.3 was pre-processed by removing punctuation and performing word tokenization using the Python "nltk.word_tokenize" module. Next, polarity intensity was obtained for each word with the use of the "senticnet" Python package. If any of the words were not recognized, the default value of 0.0 was recorded. Finally, all key statistical instruments crucial to the current project were computed for each text.

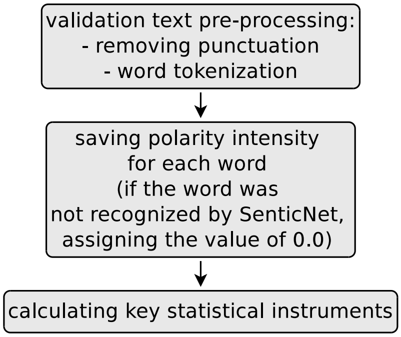

validation text pre-processing:
- removing punctuation
- word tokenization
saving polarity intensity
for each word
(if the word was
not recognized by SenticNet,
assigning the value of 0.0)
calculating key statistical instruments

Figure 1. Implementation of SenticNet in the present analysis

3.2.2 SentiStrength Lexicon

SentiStrength is a stand-alone program with a graphic user interface, but other versions of the software are also available. One is an online tool, and another is a "Java version", which is recommended for commercial use and is accessible from the command line. SentiStrength has been described and evaluated in Thelwall (2017), Thelwall et al. (2010, 2012, 2013) and Thelwall and Buckley (2013). It has also been used in numerous research projects.

The use of SentiStrength in the current study involves only the main sentiment lexicon included in the set of resources attached to the program. The lexicon is a tab separated value file with a list of English lexical items and the corresponding sentiment values on a scale of –5 to +5. The only aspect which makes the inclusion of the lexicon in the current analysis less than straightforward is the fact that a large number of the lexical items listed are inflectional or derivational bases rather than final English forms. The items meant to be the bases are marked with an asterisk at the end. For this reason, the adaptation of the lexicon for the purposes of the present study required some additional procedures (see Figure 2). The pre-processing stage of the validation materials was standard and involved removing punctuation and word tokenization using the Python "nltk.word_tokenize" module. What is different from other cases, however, is the division of the lexicon into two separate parts. All the lexical items which were "final forms" were collected in one file, and the items which were inflectional of derivational bases were saved in another file. The corresponding scoring was also saved in these files. For each text from the validation materials described in Section 3.3, all the words were searched in the first file. If any word was found, the corresponding scoring was recorded. Next, a similar search was done in the second file, but this time the results were collected not only for identical items, but also for cases in which a given word began in the same way as any of the inflectional or derivational bases. For instance, the word "abandoned" would be recognized as a possible form derived from the base "abandon*", present in the second file. Finally, the key statistical instruments defined in Section 3.1 were calculated for each text.

3.2.3 SentiWordNet (SWN)

SentiWordNet is a tool designed to be used in sentiment classification and opinion mining. It has been described in Baccianella et al. (2010), Esuli and Sebastiani (2006, 2007) and Kreutzer and Witte (2013), and applied in numerous research projects. As its name suggests, it was built using WordNet, which is a huge lexical database of English (Fellbaum 1998; Miller 1995).

SentiWordNet may be downloaded directly from "sentiwordnet.isti.cnr.it", but the version used in the present study is the module included in the Python NLTK platform (version 3.2.4) (Bird et al. 2009). Because of the rather complex structure of WordNet, application of SentiWordNet in the current analysis was more complex than the procedures used for other lexicons (see Figure 3). After importing the validation materials described in Section 3.3, punctuation was removed and word tokenization was performed using the Python "nltk.word_tokenize" module. After that, part-of-speech tagging was conducted with the use of "nltk.pos_tag". Next, function words (or "stop words", as they are referred to in NLP) were removed. The reason for this choice is the fact that, even if such items are assigned sentiment values, their interpretation depends almost entirely on context and none of the lexicons in this study takes any pragmatic aspects of texts into account. Then, lemmatization was performed using the "WordNetLemmatizer" class imported from the "nltk.stem.wordnet" module. This operation was necessary, because in the next step the SentiWordNet "senti_synset" method was used, and it recognizes correctly only uninflected forms. The method returns two separate numeric scores. Both are values between 0 and 1. The one called "PosScore" indicates the degree to which a given word is positive, and the one called "NegScore" shows the level of negative associations. Because of this rather uncommon scoring solution, the key statistical instruments defined in Section 3.1 had to be calculated differently than in other cases. Perhaps the best way to describe the exact procedure employed is to give an example. In "this intriguing girl is beautiful, but also mischievous and dangerous" some parts of the expression are positive and others are negative. The results obtained in SWN for this example are presented in Table 2.

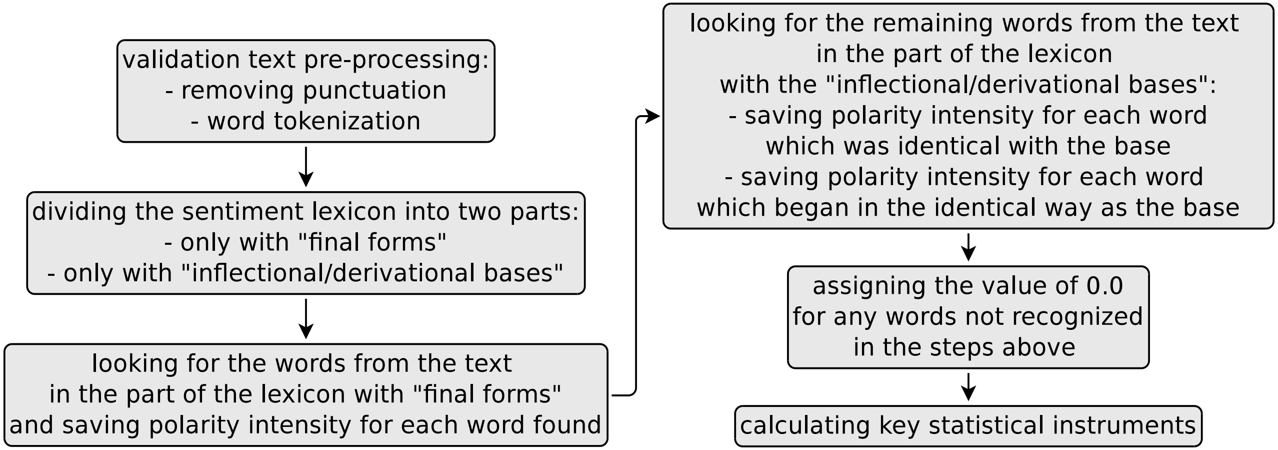

validation text pre-processing:
- removing punctuation
- word tokenization
dividing the sentiment lexicon into two parts:
- only with "final forms"
- only with "inflectional/derivational bases"
looking for the words from the text
in the part of the lexicon with "final forms"
and saving polarity intensity for each word found
looking for the remaining words from the text
in the part of the lexicon
with the "inflectional/derivational bases":
- saving polarity intensity for each word
which was identical with the base
- saving polarity intensity for each word
which began in the identical way as the base
assigning the value of 0.0
for any words not recognized
in the steps above
calculating key statistical instruments

Figure 2. Implementation of SentiStrength Lexicon in the present analysis

Table 2. Results obtained in SWN for an example text containing both positive and negative lexical items

	this	intriguing	girl	is	beautiful	but	also	mischievous	and	unpredictable
positive scores	0.0	0.5	0.0	0.0	0.75	0.0	0.0	0.0	0.0	0.0
negative scores	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.25	0.0	0.625

SUM would be measured as the difference between positive and negative scores, so (0.5 + 0.75) – (0.25 + 0.625) = 0.375. MEAN1 and MEAN2 could, however, be calculated in at least two different ways:

[1] by obtaining the mean from all positive and negative scores. MEAN1 would be calculated in the following way:
((0.5 + 0.75) – (0.25 + 0.625)) / 4 = 0.09375
MEAN2, which includes "zeroes" for words which have not been found in the lexicon, would be obtained as follows:
((0.5 + 0.75) – (0.25 + 0.625)) / 20 = 0.01875

[2] by obtaining the mean separately for positive scores and separately for negative scores, and then calculating the difference between the two results. MEAN1 would be calculated in the following way:
((0.5 + 0.75) / 2) – ((0.25 + 0.625) / 2) = 0.1875
MEAN2, similarly as before, would involve larger denominators:
((0.5 + 0.75) / 10) – ((0.25 + 0.625) / 10) = 0.0375

After performing correlation tests, it became clear that the first method was more effective. Therefore, MEAN1 and MEAN2 for SWN were calculated in the way described in [1] above.

In computing MEDIAN, only the option involving the whole set of positive and negative scores was considered (again, option [1]). This key statistical instrument requires larger datasets to indicate the middle value in a meaningful way, so there was no sense in splitting the calculations into two parts. In our example, MEDIAN = 0.125. Likewise, LAV was also obtained from the whole set of positive and negative scores. In the example under discussion LAV is 0.75.

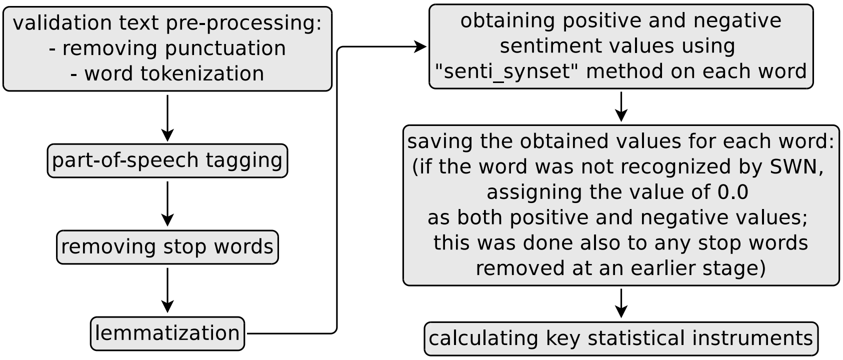

validation text pre-processing:
- removing punctuation
- word tokenization
part-of-speech tagging
removing stop words
lemmatization
obtaining positive and negative sentiment values using "senti_synset" method on each word
saving the obtained values for each word: (if the word was not recognized by SWN, assigning the value of 0.0 as both positive and negative values; this was done also to any stop words removed at an earlier stage)
calculating key statistical instruments

Figure 3. Implementation of SentiWordNet in the present analysis

3.2.4 UMass Amherst Linguistics Sentiment Corpora (UMALSC)

UMass Amherst Linguistics Sentiment Corpora (Constant et al. 2009; Potts – Schwarz 2008) is a collection of n-gram counts extracted from a large number of online product reviews in four languages: Chinese, English, German, and Japanese. The part which is of interest to the present study are the eight datasets for the English language. Four of them contain statistics on bigrams, and the other four focus on unigrams. Although both unigrams and bigrams could be used in this study, for the sake of simplicity only unigrams were utilized. The four files were created on the basis of the following sources: 1) English Amazon book reviews; 2) English Amazon book summaries; 3) English Tripadvisor.com reviews; and 4) English Tripadvisor.com summaries.

Table 3. An example of the structure of UMALSC unigrams counts

Token	Rating	TokenCount	RatingWideCount
absurd	1	35	570687
absurd	2	27	512643
absurd	3	14	767958
absurd	4	20	1513776
absurd	5	48	4769921
abundance	1	8	570687
abundance	2	7	512643
abundance	3	22	767958
abundance	4	43	1513776
abundance	5	109	4769921

An example of the structure of the files is shown in Table 3. For each word type, token counts are provided for 5 ratings. The ratigs are on a gradable scale of 1 (very negative) to 5 (very positive). Additionally, the total token count for each rating is also provided. Such raw data needed to be processed in order to obtain a single sentiment score for each word type. The solution chosen involved two stages. Firstly, the four datasets were concatenated into one. Each word type present in any of the four files was searched for in the other three files. If it was present only in this dataset, the token counts were just copied to the concatenated file. However, if a given word type was found in more than one dataset, its token counts were added and the sum was saved in the concatenated file instead. Secondly, a single, unidimensional measure of sentiment for each word type was calculated using the formula presented below. x represents "token count" for a given rating, and w is the weight assigned to each rating ($w_1 = –2$, $w_2 = –1$, $w_3 = 0$, $w_4 = 1$, $w_5 = 2$).

$$\frac{\sum_{i=1}^{5} x_i \, w_i}{5 \sum_{i=1}^{5} x_i}$$

The resulting sentiment lexicon was implemented in the present analysis in the same way as the Vader Lexicon described below (see Section 3.2.5 and Figure 4).

3.2.5 Vader Lexicon (VL)

The lexicon described in this section is included in the sentiment analysis tool known as "Valence Aware Dictionary and sEntiment Reasoner" or VADER (Hutto – Gilbert 2014). The tool is available as a Python library and it involves both a lexicon and a rule-based sentiment analysis algorithm. Nevertheless, this study focuses only on the former component, which will be referred to as "Vader Lexicon" or VL. The lexicon is a tab delimited file. It provides sentiment ratings on a scale of –4 (very negative) to +4 (very positive) for over 7000 word types created on the basis of ratings provided by multiple independent human judges. The lexicon is especially attuned to social media contexts but may be useful for sentiment analysis in other domains.

The way in which Vader Lexicon has been applied in the present study is summarised in Figure 4. In the validation texts discussed in Section 3.3 all the punctuation was removed and word tokenization was performed with the use of the “nltk.word_tokenize” module. Next, part-of-speech tagging was conducted using “nltk.pos_tag”. Then, function words were removed from the texts. Since VL does not include such lexical items, this step was optional and it was performed solely for increasing the speed of processing. After that, the “WordNetLemmatizer” class imported from the “nltk.stem.wordnet” module was used for lemmatization. Sentiment values were obtained in a three-step procedure. If a given word type was found in the lexicon, the value was assigned to it directly. If the word type was not found, however, the corresponding lemma was searched for in VL. This step maximized the number of lexical items scored and potentially improved the general performance of the lexicon. Finally, if the word type was not found in any of the two previous steps, the default value of 0.0 was assigned to it. The same was done to any function words removed at an earlier stage.

The key statistical instruments for each text were calculated in the standard manner described in Section 3.1.

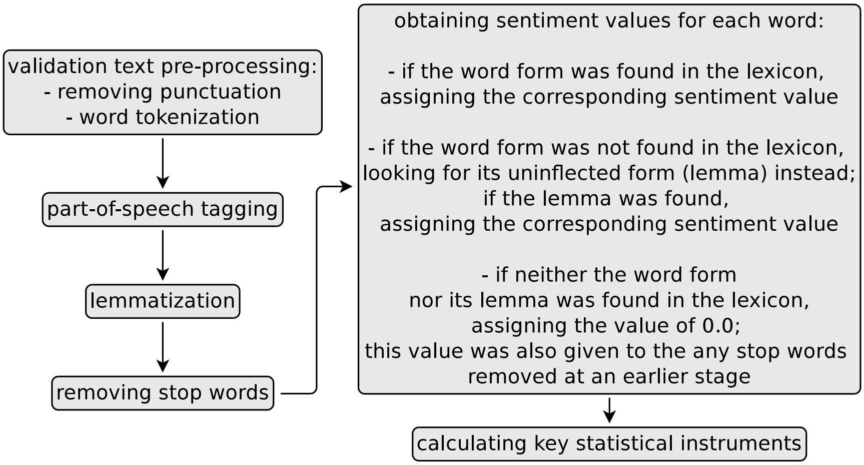

validation text pre-processing:
- removing punctuation
- word tokenization
part-of-speech tagging
lemmatization
removing stop words
obtaining sentiment values for each word:
- if the word form was found in the lexicon, assigning the corresponding sentiment value
- if the word form was not found in the lexicon, looking for its uninflected form (lemma) instead; if the lemma was found, assigning the corresponding sentiment value
- if neither the word form nor its lemma was found in the lexicon, assigning the value of 0.0; this value was also given to the any stop words removed at an earlier stage
calculating key statistical instruments

Figure 4. Implementation of Vader Lexicon in the present analysis

3.2.6 Sentiment and Emotion Lexicons created by the National Research Council Canada (NRC)

The tools offered by the National Research Council Canada include a variety of different sentiment and emotion lexicons. Nine of these lexicons involve numerical scoring for the dichotomy “positive” vs. “negative” and are appropriate for current purposes. They are briefly described below.

[1] Sentiment Composition Lexicon for Negators, Modals, and Degree Adverbs (SCL-NMA) has been described in Kiritchenko and Mohammad (2016b). The lexicon comprises 1621 single words and 1586 phrases. The phrases were formed by combining single words with an auxiliary verb, a degree adverb, a negator, or a combination of those. Each single word or multiple-word phrase has been given a sentiment score on a scale of –1 (very negative) to +1 (very positive). The scores were obtained through crowdsourcing.

[2] SemEval-2015 English Twitter Lexicon (ETL) has been discussed in Kiritchenko et al. (2014). It contains 1515 single words and two-word phrases. All of them have been taken from English Twitter. The two-word expressions are composed of words proceeded by negators. Each single word and two-word phrase has been given a sentiment value on a scale of –1 (very negative) to +1 (very positive). As in the previous case, the scores were obtained through crowdsourcing.

[3] Sentiment Composition Lexicon for Opposing Polarity Phrases (SCL-OPP) (Kiritchenko – Mohammad 2016a, 2016c) consists of 1178 unigrams, bigrams and trigrams which were taken from tweets. The two-word and three-word phrases contain at least one positive word and at least one negative word. Again, the sentiment scoring involves a scale of –1 to +1 and it was obtained through crowdsourcing.

[4] NRC Hashtag Sentiment Lexicon (HSL) (Kiritchenko et al. 2014; Mohammad et al. 2013; Zhu et al. 2014) consists of 50836 unigrams and 245920 bigrams. A file with pairs of unigrams and bigrams is also available, but it has not been used in this paper. The unigrams and bigrams were automatically generated from 775000 tweets with sentiment-word hashtags. Each unigram and bigram has been assigned a sentiment value using the algorithm described in Kiritchenko et al. (2014, p. 732). Most of the scores are between –2 and +3, but in extreme cases they reach values above 8.

[5] Hashtag Affirmative Context Sentiment Lexicon and Hashtag Negated Context Sentiment Lexicon (HSL-AFF-NEG) (Kiritchenko et al. 2014; Mohammad et al. 2013; Zhu et al. 2014) contains 43904 unigrams and 174904 bigrams. For some unigrams and bigrams it was indicated that they were taken from negated contexts. For the purposes of the current study, all such cases were removed from the lexicon. The

unigrams and bigrams were generated from the source used in HSL and sentiment scores were assigned using the same method as in the previous case.

[6] Emoticon Lexicon (EL) (Kiritchenko et al. 2014; Mohammad et al. 2013; Zhu et al. 2014) was automatically generated from 1.6 million tweets with emoticons. As in the case of HRC, the lexicon is divided into unigrams (62447 words), bigrams (641737 two-word phrases) and pairs of unigrams and bigrams, but again, the latter file has not been utilized in this study. Sentiment values were assigned using identical methods as in the previous two cases.

[7] Emoticon Affirmative Context Lexicon and Emoticon Negated Context Lexicon (EL-AFF-NEG) (Kiritchenko et al. 2014; Mohammad et al. 2013; Zhu et al. 2014) is based on the same materials as EL. Moreover, the same methods were used in assigning sentiment scores. The lexicon comprises 55054 unigrams and 262142 bigrams. As in the case of HSL-AFF-NEG, some items were marked for having been taken from negative contexts. For the current purposes, such unigrams and bigrams were removed from the lexicon.

[8] Yelp Restaurant Sentiment Lexicon (YRSL) (Kiritchenko et al. 2014) contains 39232 unigrams and 268303 bigrams and was automatically generated from customer reviews from the Yelp Phoenix Academic Dataset available at "http://www.yelp.com/dataset_challenge". Sentiment scores were assigned automatically using the same methods as in the previous four cases. Again, some examples were removed from this lexicon because the context in which they were originally used involved negation.

[9] Amazon Laptop Sentiment Lexicon (ALSL) (Kiritchenko et al. 2014) was generated from reviews on laptops and notebooks collected from "Amazon.com". The lexicon includes 26561 unigrams and 149118 bigrams. Sentiment scores were assigned in the same manner as in the previous five cases. Also, the unigrams and bigrams which were marked for coming from negated contexts were removed as in the case of HSL-AFF-NEG, EL-AFF-NEG and YRSL.

A summary of the processing used in preparing NRC lexicons for the current study is presented in Figure 5. The first two steps have been discussed in the individual descriptions above. The third stage, however, requires further explanation. After removing punctuation, some expressions were duplicated and a Python script was written to merge them into one. The resulting sentiment scoring was the mean of the scores for all the instances of the duplicated expression. Because of the large size of NRC lexicons, this part of processing was performed at the Mathematical Modelling Laboratory at Jan Kochanowski University in Kielce, Poland.

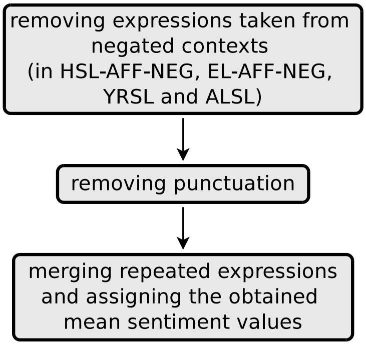

removing expressions taken from negated contexts (in HSL-AFF-NEG, EL-AFF-NEG, YRSL and ALSL)
removing punctuation
merging repeated expressions and assigning the obtained mean sentiment values

Figure 5. Processing involved in preparing NRC lexicons for the current project

On the basis of the NRC tools described above, a new lexicon was created. It will be referred to as NRC Combined or “NRCC”. It is an aggregate of all NRC lexicons. The process of generating it involved copying all the expressions and their sentiment scores into one file. After that, duplicated expressions were merged. The assigned sentiment scores were the means calculated from all the instances of a given, duplicated expression. The task was computationally demanding and, again, the processing was performed at the Mathematical Modelling Laboratory at Jan Kochanowski University in Kielce, Poland.

Figure 6 shows the implementation of all NRC lexicons in the present analysis. After standard preprocessing involving punctuation removal and word tokenization, part-of-speech tagging was performed on each validation text using the Python “nltk.pos_tag” package. Next, lemmatization was conducted with the “WordNetLemmatizer” class imported from the “nltk.stem.wordnet” module. Sentiment values were obtained in a procedure more complex than those of the cases described previously. NRC lexicons prepared for the present analysis contained unigrams, bigrams and trigrams. In each validation text, trigrams were searched first. If any were found, the corresponding sentiment values were recorded and the trigrams were removed from the text. Next, bigrams were searched. If any were found, the sentiment values were saved and the bigrams were removed from the text. The same procedure was repeated for unigrams. Additionally, the lemmas of the remaining words were searched in the unigrams part of a given dictionary and if any were found, the corresponding sentiment values were recorded. Finally, the value of 0.0 was assigned to any remaining lexical items and the statistical instruments under analysis were calculated in the standard way described in Section 3.1.

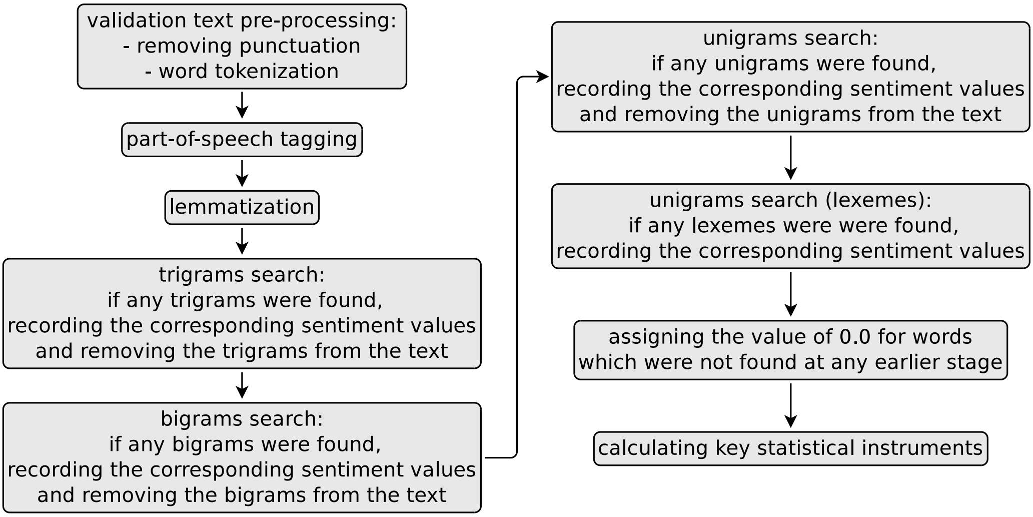

validation text pre-processing:
- removing punctuation
- word tokenization
part-of-speech tagging
lemmatization
trigrams search:
if any trigrams were found,
recording the corresponding sentiment values
and removing the trigrams from the text
bigrams search:
if any bigrams were found,
recording the corresponding sentiment values
and removing the bigrams from the text
unigrams search:
if any unigrams were found,
recording the corresponding sentiment values
and removing the unigrams from the text
unigrams search (lexemes):
if any lexemes were were found,
recording the corresponding sentiment values
assigning the value of 0.0 for words
which were not found at any earlier stage
calculating key statistical instruments

Figure 6. Implementation of NRC lexicons in the present analysis

3.3 Validation materials

The validation materials used in this study were taken from two independent projects on sentiment analysis. Four of the datasets come from the VADER project described in Hutto and Gilbert (2014) (see also Section 3.2.5). The other six were downloaded from the official website of SentiStrength and they are characterized in Thelwall et al. (2012) (see also Section 3.2.2).

A summary of the validation materials is presented in Table 4. It must be stressed, however, that this summary provides statistics on the way the data were used in this study, rather than on their original characteristics. The author could not get access to some parts of the raw materials and some examples were excluded in a few datasets. (Similar problems were encountered in previous studies involving the currently used test data, e.g. Ribeiro et al. 2016).

The materials differ in terms of the number of texts. The dataset with the smallest number of snippets is "BBC forum posts" (693 texts), the one with the largest number is "Rotten Tomatoes movie reviews" (over 10000 texts) and the average number of texts for all 10 datasets is 3349.6. A quick glance at Table 4, however, suggests, that it is also necessary to take into account other aspects of the excepts, such as the average number of words in each fragment in a given dataset[1]. For instance, the mean text length in "Rotten Tomatoes movie reviews" (18.83 words) is much shorter than the average text length in "BBC forum posts" (60.79). Therefore, a statistical instrument which considers the overall number of words of each dataset should be used. Indeed, "Rotten Tomatoes movie reviews" is the largest of the test datasets used in this study. It contains almost 200000 words. The smallest, on the other hand, is "MySpace comments", with 20001 words. The average number of words for all 10 datasets is 66304.3 and the sum of all the words in the data amounts to 663043.

Not only are the validation materials used in this paper extensive, but they represent different types of social Internet environments. The materials obtained from the SentiStrength website concentrate on various comments and posts (Thelwall et al. 2012). "BBC forum posts" involve discussions about various serious topics, "Digg posts" represent news commentaries, "Runners World forum posts" include messages exchanged by a common-interest group, "Twitter posts 2" are public blog broadcasts and "YouTube comments" represent comments on resources available at "youtube.com." The test materials downloaded from Vader's GitHub repository, on the other hand, focus on reviews ("Amazon product reviews", "Rotten Tomatoes movie reviews") and opinion news articles ("New York Times opinion editorials"). "Twitter posts 1" are similar to "Twitter posts 2", but according to the description on Vader's website, they are "tweet-like" texts "inspired" by tweets rather than unaltered messages obtained directly from "twitter.com".

Each text in the 10 datasets was rated for sentiment value by (a) human participant(s). What is most crucial, however, is the fact that the ratings are not polarity-based, but valence-based. Instead of classifying the texts as positive or negative, a gradable scale was used. In the case of the Vader datasets, the scale was from –4 (extremely negative) to +4 (extremely

[1] The number of words in each dataset was calculated in Python using the "nltk. tokenize" package. Emoticons and other symbols which are not part of the Roman alphabet were excluded.

Table 4. Statistics on the validation materials

dataset	abbreviation	source	type of texts	number of texts	mean text length (number of words)	sd of text length (number of words)	number of words
Amazon product reviews	amazon_reviews	Vader	customer reviews	2693	15.96	10.37	42969
Rotten Tomatoes movie reviews	movie_reviews	Vader	movie reviews	10605	18.83	8.69	199726
New York Times opinion editorials	nyt_editorial	Vader	opinion editorials	5181	17.42	8.71	90322
Twitter posts 1	tweets	Vader	tweets	4198	13.41	6.69	56308
BBC forum posts	bbc	SentiStrength	social media comments	693	60.79	72.09	42311
Digg posts	digg	SentiStrength	social media comments	1077	31.45	44.23	33873
MySpace comments	myspace	SentiStrength	social media comments	1041	19.21	25.10	20001
Runners World forum posts	rw	SentiStrength	social media comments	1046	63.61	68.44	66545
Twitter posts 2	twitter	SentiStrength	social media comments	3555	14.94	6.39	53142
YouTube comments	youtube	SentiStrength	social media comments	3407	16.98	17.83	57846
all materials mean				3349.6	27.26	26.854	66304.3
all materials sum				33496			663043

positive). Any values in between represented more moderate attitudes, with 0 indicating neutrality. Similarly, SentiStrength materials were graded on a scale of 1 to 5, but separately for positive and negative emotions. For instance, a text regarded as extremely positive would be given 5 on the "positive emotion scale" and 1 on the "negative emotion scale". In the present study, "negative scores" were deducted from "positive scores" and the resulting value was assumed to represent the sentiment value of a given text. For example, if the score on the "positive emotion scale" was 5 and on the "negative emotion scale" was 1, the score used for the present paper was 4. Consequently, the range of the scoring was from –4 to +4, just as in the previous case.

3.4 Statistical tests

The analysis described in Section 4 involves performing tests for association between paired samples, using Pearson's product moment correlation coefficient. The independent variables are datasets with the key statistical instruments obtained for each text in the validation materials according to each lexicon. For instance, for the entire "Amazon product reviews" collection there are as many as 75 such datasets (5 types of key statistical instruments × 15 lexicons). In each case, the dependent variable is the corresponding human scoring. The correlation coefficients obtained are further tested with the use of ANOVA and Tukey HSD post-hoc analysis. The choice of these parametric methods is based on the observation that both the normality condition and the equal variance condition are not severely violated (see Figure 7).

It is worth noting that the independent variables involve numeric results on different scales. This stems from the fact that the sentiment lexicons themselves use different scoring ranges. Nevertheless, no attempt has been made to normalize the data since it is not really a problem for the correlation tests as long as the scale is the same for all the data in a given set. The resulting correlation coefficients will be the same, no matter what the range of the scoring scale is.

All the statistical tests were performed using R (R Development Core Team 2013).

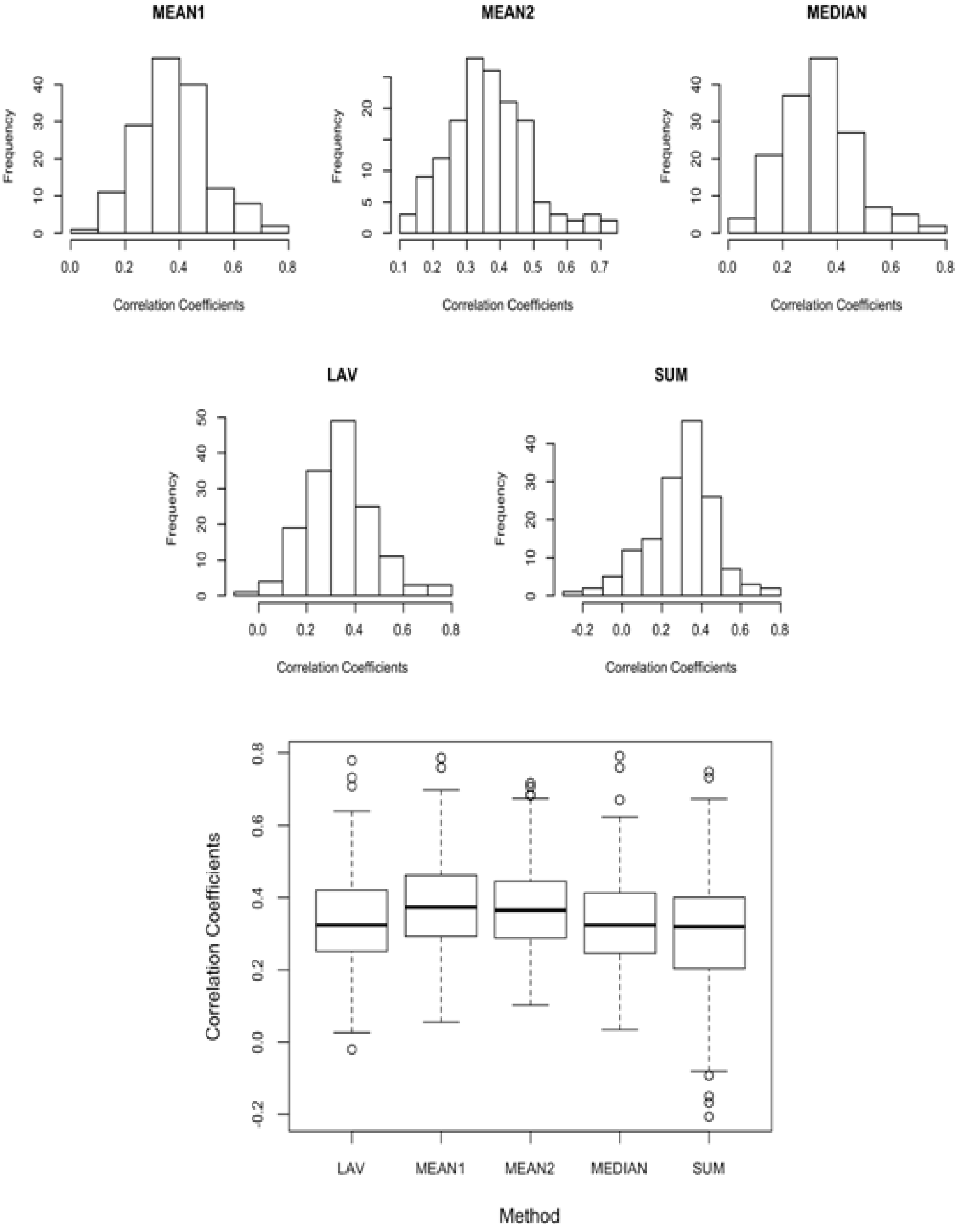


Figure 7. Histogram and boxplot of the correlation coefficients obtained for data including negation (similar patters were observed for data excluding negation)

4. Results

The results are summarised separately for tests performed on entire validation datasets (Section 4.1) and tests conducted on the validation materials divided into smaller samples (Section 4.2).

4.1 Results based on entire validation sets

The general results of the analysis performed in this study are presented in Table 5. They are based on 750 correlation tests (5 types of key statistical instruments × 15 lexicons × 10 validation sets). The values in the column "average correlation for all texts" were obtained by averaging correlation coefficients for all validation sets with all the texts they include. It is immediately visible that the method with the highest mean correlation is MEAN1 (0.38189). The second most successful measure is MEAN2, with the result of 0.36916, which is 0.01273 less than in the case of MEAN1. MEDIAN and LAV performed on a similar level. The mean correlation for both is around 0.33. Finally, the method with the lowest result is SUM. Its mean correlation is only 0.29631, which is almost 0.1 less than the mean correlation obtained for MEAN1.

Table 5. Average correlation for key statistical instruments based on entire validation sets

method	average correlation for all texts	average correlation excluding texts with negation
MEAN1	0.38189	0.40189
MEAN2	0.36916	0.38673
MEDIAN	0.33607	0.35716
LAV	0.33121	0.35905
SUM	0.29631	0.34213

An ANOVA for the data discussed above was performed. It revealed a statistically significant difference between at least two groups ($F(4,745) = 8.654$, $p < 0.0001$), so a Tukey HSD post-hoc analysis was also conducted. The pairs whose comparison yielded p-values below the alpha level of 0.05 are listed below:

- MEAN1 – MEDIAN ($p = 0.0392$)
- MEAN1 – LAV ($p = 0.016$)
- MEAN1 – SUM ($p < 0.0001$)
- MEAN2 – SUM ($p < 0.0001$)

These results clearly indicate that MEAN1 is, in fact, the most effective method of those under investigation. The only possible exception is MEAN2, whose lower ranking has not been statistically substantiated. Besides the four pairs above, no other comparison indicated that the differences were statistically significant. One of the confounding factors which might explain this is the fact that some of the texts in the validation tests involve negation, which is a broadly discussed issue in sentiment analysis (see the last paragraph in Section 1). Many tools, such as Vader or Sentistrength, are designed to cope with it, but the simplistic processing used in calculating the five key statistical instruments in this study does not deal with this problem at all. Therefore, a more appropriate solution is to use validation materials without texts involving negation.

A script was written in Python and all examples with negation were removed. Next, another series of correlation tests was conducted. The results obtained are summarised in Table 5 in the column "average correlation excluding texts with negation". They are based on exactly the same number of tests (5 types of key statistical instruments × 15 lexicons × 10 validation sets = 750 correlation tests), but this time each validation set is shorter and does not include any negated sentences. The average correlation coefficients obtained are higher by approximately 0.02, with the exception of the result for SUM, which is higher by almost 0.05. The relative ranking of the methods, however, does not change in any significant way. Again, the statistical instrument which produces the best results is MEAN1, with MEAN2 close behind, followed by MEDIAN and LAV. The result for SUM is still the worst, although the gap between it and the other methods is smaller than in the previous analysis. An ANOVA performed on these data indicated statistically significant difference between at least two groups ($F(4,745) = 4.494$, $p = 0.0014$), so a Tukey HSD post-hoc analysis was conducted once again. The comparison of each possible pair yielded results very similar to those of the previously performed Tukey HSD. In fact, the p-values obtained are slightly higher. The ones still indicating statistical significance are listed below:

- MEAN1 – MEDIAN (p = 0.0.0463)
- MEAN1 – SUM (p = 0.0022)
- MEAN2 – SUM (p = 0.0473)

The result for the pair MEAN1 – LAV is marginally significant (p = 0.0634). All the rest of the comparisons indicate that the differences cannot be statistically confirmed.

4.2 Results based on validation sets divided into smaller samples

A factor which is important in the tests performed thus far is the size of samples. In the previous calculations, the mean correlation coefficient for each method has been computed on the basis of 150 correlation tests (15 lexicons × 10 validation sets). The ANOVA and Tukey HSD performed later did not take into account the actual size of the samples on which the correlation tests were performed. It is, however, possible to divide the 10 validation datasets into smaller sets. In this way the number of the actual correlation tests could be increased substantially.

Table 6. Division of validation materials into smaller samples

dataset	including negation		excluding negation	
	number of texts	number of samples	number of texts	number of samples
Amazon product reviews	2693	27	2044	21
Rotten Tomatoes movie reviews	10605	106	8071	81
New York Times opinion editorials	5181	52	4310	43
Twitter posts 1	4198	42	3056	31
BBC forum posts	693	7	375	4
Digg posts	1077	11	704	7
MySpace comments	1041	11	881	9
Runners World forum posts	1046	11	655	7
Twitter posts 2	3555	36	3056	31
YouTube comments	3407	34	2844	29
sum	33496	337	25996	263

Statistics coursebooks (e.g. Rumsey 2003) usually suggest that the minimum sample size for obtaining reliable results is around 30. Since the validation datasets used in this study are large, samples of 100 were eventually chosen, if the final group of texts totalled at least 30, it was included in the analysis. The way in which each validation dataset was divided into smaller sets is presented in Table 6. For instance, "Amazon product reviews" contains 2693 texts. Consequently, 26 samples of 100 texts were obtained plus one final sample with 93 texts. Since this final sample is larger than the minimum of 30, it has been included in the analysis, so the ultimate number of samples for this validation dataset is 27. However, in the dataset "Rotten Tomatoes movie reviews", the last sample was ignored because it contains only 5 texts.

After dividing the validation materials into smaller samples, as many as 25275 correlation tests were performed (5 types of key statistical instruments × 15 lexicons × 337 validation sets). This time, each mean correlation coefficient for each key statistical instrument was calculated on the basis of 5055 measurements (15 lexicons × 337 validation sets). The results obtained are summarised in Table 7 in the column "average correlation for all texts". The ranking of the methods is identical to the hierarchy observed before. MEAN1 is the most efficient statistical instrument, closely followed by MEAN2. On this occasion, however, the difference is smaller than before (only about 0.006). The results for the other three methods are very similar to each other, but clearly lower than the average correlations obtained for MEAN1 and MEAN2. An ANOVA performed on these data revealed a statistically significant difference between at least two groups ($F(4,25270) = 106.4$, $p < 0.0001$), so a Tukey HSD post-hoc analysis was conducted. Here, the majority of the differences between the results are statistically significant, with the exception of the four pairs listed below.

Table 7. Average correlation for key statistical instruments based on validation sets divided into smaller samples

method	average correlation for all texts	average correlation excluding texts with negation
MEAN1	0.35900	0.38138
MEAN2	0.35296	0.37338
MEDIAN	0.31448	0.33712
LAV	0.30926	0.33638
SUM	0.30954	0.33916

- MEAN1 – MEAN2 (p = 0.4713)
- MEDIAN – LAV (p = 0.6338)
- MEDIAN – SUM (p = 0.6877)
- LAV – SUM (p = 0.9999)

These analyses show that both MEAN1 and MEAN2 are more effective statistical instruments than MEDIAN, LAV and SUM. No other differences, however, have been confirmed. The same conclusions can be drawn from the analysis excluding texts involving negation. Although the average correlation coefficients summarised in Table 7 are higher by over 0.02, the ranking of the methods, as well as the relative differences between the way they performed, is identical to the analysis with "all texts". Indeed, a Tukey HSD post-hoc analysis has revealed that only differences between the same four pairs (MEAN1 – MEAN2, MEDIAN – LAV, MEDIAN – SUM, LAV – SUM) cannot be confirmed statistically.

5. Conclusion

Lexicon-based tools for sentiment analysis frequently involve complex rule-based sentiment analysis algorithms. These algorithms aim at compensating for various linguistic phenomena, such as negation and intensification. Ultimately, they summarise the analysis performed by providing either a specific category (e.g. "positive" or "negative") or a numeric polarity score.

In the present study, an investigation was made into key statistical instruments that may be used to obtain such final results. The data analysed indicate that, other things being equal, the mean is more effective than the median, the largest absolute value, or the sum in obtaining numeric polarity scores similar to the scores provided by human participants. This conclusion may be useful in improving software packages performing lexicon-based sentiment analysis. If there is no compelling reason to use other statistical instruments in the calculation of the final score, the mean is the best option. Such a decision may also be made by the end users of tools which offer a variety of options for calculating final sentiment scores (e.g. SentiStrength).

As far as the exact way in which the mean should be calculated, no definitive answer can be offered. Two different methods were tested, one excluding lexical items not found in a given sentiment lexicon and the other including such items. Neither of the two methods was clearly more efficient than the other.

REFERENCES

Abbasi, A. – A. Hassan – M. Dhar

2014 "Benchmarking Twitter sentiment analysis tools." In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, 26-31.

Aue, A. – M. Gamon
2005 "Customizing sentiment classifiers to new domains: A case study". In: *Proceedings of Recent Advances in Natural Language Processing (RANLP)*. Borovets, Bulgaria, September 17-19, 2005.

Baccianella, S. – A. Esuli – F. Sebastiani
2010 "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." In: *Proceedings of the 7th International Conference on Language Resources and Evaluation* (*LREC),* 2200-2204.

Bird, S. – E. Klein – E. Loper
2009 *Natural Language Processing with Python*. Beijing: O'Reilly.

Borth, D. et al.
2013 "Large-scale visual sentiment ontology and detectors using adjective noun pairs". In: A. Jaimes (ed.) *Proceedings of the 21st ACM International Conference on Multimedia*. New York: ACM, 223-232.

Cambria, E. et al.
2016 "SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives". In: *Proceeding of COLING 2016, The 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan, December 11-17 2016, 2666-2677.

Choi, Y. – C. Cardie
2008 "Learning with compositional semantics as structural inference for subsentential sentiment analysis". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, October 2008,* 793-801.

Constant, N. et al.
2009 "The pragmatics of expressive content: Evidence from large corpora", *Sprache und Datenverarbeitung 33* (1-2), 5-21.

Diniz, J.P. et al.
2016 "iFeel 2.0: A multilingual benchmarking system for sentence-level sentiment analysis". In: *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*. Association for the Advancement of Artificial Intelligence.

Esuli, A. – F. Sebastiani
2006 "SENTIWORDNET: A publicly available lexical resource for opinion mining". In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. May 2006, Genoa.
2007 "SENTIWORDNET: A high-coverage lexical resource for opinion mining". *Technical Report 2007-TR-02*. Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche. Pisa, Italy.

Fellbaum, C.
1998 *WordNet: An Electronic Lexical Database*. Cambridge, Mass.; London: MIT Press.

Gonçalves, P. et al.

2013 "Comparing and combining sentiment analysis methods". In: *Proceedings of the First ACM Conference on Online Social Networks*, 27-38.

Hutto, C.J. – E. Gilbert

2014 "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Proceedings of The Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14)*, 216-255.

Kennedy, A. – D. Inkpen

2006 "Sentiment classification of movie reviews using contextual valence shifters", *Computational intelligence* 22 (2), 110-125.

Kiritchenko, S. – S. Mohammad

2016a "Happy accident: A sentiment composition lexicon for opposing polarity phrases". In: *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*. Portorož, Slovenia.

2016b "The effect of negators, modals, and degree adverbs on sentiment composition". In: *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*. San Diego, California.

2016c "Sentiment composition of words with opposing polarities". In: *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL).* San Diego, California, 1102-1108.

Kiritchenko, S. – X. Zhu – S.M. Mohammad

2014 "Sentiment analysis of short informal texts", *Journal of Artificial Intelligence Research 50*, 723-762.

Kreutzer, J. – N. Witte

2013 *Opinion mining using SentiWordNet*. Semantic Analysis, HT 2013/14. Uppsala University. http://santini.se/teaching/sais/Ass1_Essays_FinalVersion/Kreutzer_Julia_AND_Witte_Neele_SentiWordNet_Neele+Julia_finalversion.pdf, accessed December 2021

Liu, B.

2012 "Sentiment analysis and opinion mining", *Synthesis Lectures on Human Language Technologies 5* (1), 1-167.

Liu, B. – L. Zhang

2012 "A survey of opinion mining and sentiment analysis". In: C.C. Aggarwal – C. Zhai (eds.) *Mining Text Data.* Springer, 415-463.

Miller, G. A.

1995 "WordNet: A lexical database for English", *Communications of the ACM* 38 (11), 39-41.

Mohammad, S.M. – S. Kiritchenko – X. Zhu

2013 "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets". In: *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, June 2013, Atlanta, USA.

Mohammad, S.M. – P.D. Turney

2010 "Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon". In: *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Los Angeles, California: Association for Computational Linguistics, 26-34.

2013 "Crowdsourcing a word–emotion association lexicon", *Computational Intelligence* 29 (3), 436-465.

Nielsen, F.Å.

2011 "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs". In: *Proceedings of the ESWC2011 Workshop on "Making Sense of Microposts": Big things come in small packages (2011)*, 93-98.

Pang, B. – L. Lee

2008 "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval* 2 (1-2), 1-135.

Potts, C. – F. Schwarz

2008 "Exclamatives and heightened emotion: Extracting pragmatic generalizations from large corpora", *Ms., UMass Amherst*, 1-29.

R Development Core Team

2013 *R: A Language and Environment for Statistical Computing*. [computer software]. Version 3.0.3.

Ribeiro, F.N. et al.

2016 "SentiBench – A benchmark comparison of state-of-the-practice sentiment analysis methods", *EPJ Data Science 5* (23), 1-29.

Rumsey, D.J.

2003 *Statistics for Dummies*. Hoboken, N.J.: Wiley Publishing.

Taboada, M. et al.

2011 "Lexicon-based methods for sentiment analysis", *Computational Linguistics 37* (2), 267-307.

Thelwall, M.

2017 "Heart and soul: Sentiment strength detection in the social web with SentiStrength". In: J. Holyst (ed.) *Cyberemotions: Collective Emotions in Cyberspace.* Berlin: Springer, 119-134.

Thelwall, M. – K. Buckley

2013 "Topic-based sentiment analysis for the social web: The role of mood and issue-related words", *Journal of the Association for Information Science and Technology 64* (8), 1608-1617.

Thelwall, M. – K. Buckley – G. Paltoglou

2012 "Sentiment strength detection for the social web", *Journal of the Association for Information Science and Technology* 63 (1), 163-173.

Thelwall, M. et al.

2010 "Sentiment strength detection in short informal text", *Journal of the American Society for Information Science and Technology 61* (12), 2544-2558.

Thelwall, M. et al.
2013 "Damping sentiment analysis in online communication: Discussions, monologs and dialogs". In: A. Gelbukh (ed.) *Computational Linguisticsand Intelligent Text Processing, 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II.* Springer, 1-12.

Zhu, X. – S. Kiritchenko – S. Mohammad
2014 "NRC-Canada-2014: Recent improvements in the sentiment analysis of tweets". In: P. Nakov – T. Zesch (eds.) *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics, 443-447.

Address: ŁUKASZ STOLARSKI, Uniwersytet Jana Kochanowskiego w Kielcach, Instytut Literaturoznawstwa i Językoznawstwa, ul. Uniwersytecka 17, 25-406 Kielce, Poland.
ORCID code: https://orcid.org/0000-0002-2668-5509