LiSpeTT

Linguaggi specialistici e Traduzione Tecnica

Volume 1

Representing and Redefining Specialised Knowledge: Medical Discourse

edited by
Anthony Baldry
Francesca Bianchi
Anna Loiacono

UNIVERSITÀ DEL SALENTO

2019

LiSpeTT

Volume 1

"Linguaggi specialistici e Traduzione Tecnica" (LiSpeTT) è una collana del Dipartimento di Studi Umanistici dell'Università del Salento. La Collana pubblica lavori originali di carattere scientifico (monografie, volumi collettanei, atti di convegno e seminari), frutto delle ricerche svolte sia da studiosi di questo ateneo sia da studiosi di altri atenei nazionali e internazionali rigorosamente nell'ambito delle due tematiche specificate, ossia linguaggi specialistici e traduzione tecnica, che possono anche essere considerate singolarmente. I volumi proposti per una pubblicazione nella collana saranno sottoposti a controllo antiplagio e valutazione attraverso il procedimento del doppio referaggio anonimo (double-blind peer-review process). Questa collana accoglie opere redatte in italiano, inglese, francese, tedesco, spagnolo, portoghese e russo.

"Linguaggi specialistici e Traduzione Tecnica" (LiSpeTT – Specialised Languages and Translation) is a journal series of the Department of Humanities of the University of Salento. The series publishes original scientific works (monographs, multi-authored edited thematic collections, workshop and conference proceedings) carried out by scholars belonging to the University of Salento as well as by national and international scholars. The works published in the series must revolve around the following two topics: specialized languages and/or technical translation. All eligible proposals will be screened for plagiarism and will go through a double-blind peer-review process. The series publishes works written in English, French, German, Spanish, Portuguese, and Russian.

OFFICES:
Dipartimento di Studi Umanistici
73100 LECCE, via Taranto, 35
tel. +39-0832-294208/06, fax +39-0832-249427

http://siba.unisalento.it
ISBN 978-88-8305-153-1
ISSN XXXX-XXXX
http://siba-ese.unisalento.it

Coordinamento SIBA UNIVERSITÀ DEL SALENTO

Table of contents

CHAPTERS

Chapters

PREFACE TO THIS VOLUME

ANTHONY BALDRY, FRANCESCA BIANCHI, ANNA LOIACONO

1. Our thanks to contributors and reviewers

As editors of *Representing and Redefining Specialised Knowledge: Medical discourse*, our first duty is to thank both our reviewers and contributors for the hard work they have put into the production of this volume. Each of the papers selected for publication underwent two rigorous blind peer reviews and further revision in the light of various editorial suggestions. The best way of expressing our thanks is to ensure the widest possible readership for this volume. The fact that this is the first volume in the *LISPET - Linguaggi Specialistici e Traduzione Tecnica* series and can be downloaded free of charge from the *ESE - Salento University Publishing* website (*http://siba-ese.unisalento.it/*) is a first step in this direction.

2. Our message to university students

Our second duty relates to university students, all of whom engage with discourse in English in relation to specialised knowledge. We invite you to download this volume as we believe you will benefit from thinking about how specialised knowledge is represented, defined and accessed in today's society. We also advise you to think about how digital corpora can help you master specialised discourse in English. For example, you will sometimes want to check up on specific structures in English, which is where even a rudimentary knowledge of what corpora are and how they represent scientific discourse can be very useful. Like dictionaries and other online tools, online corpora can tell you whether the expressions you would like to use are correctly formed and whether those you come across in lectures or in textbooks are frequently used forms or, on the contrary, rather rare constructions. But, unlike other tools, online corpora go much further in providing you with a more complete picture of the meanings of the expressions you encounter and how best to use them in your own discourse. For example, they are a good guide to those contexts where the expressions you want to check up on are typically used. They also help you identify those

Doi Code: 10.1285/i9788883051531p7

contexts where these expressions are seldom or never used. So regardless of the degree course you are following, we encourage you to download this volume and to reflect just a little on how online corpora and corpus studies can help you.

For those students whose degree courses require them to constantly engage with specialised knowledge of English in their daily studies, and who need to understand the nature of scientific discourse more fully, may we recommend the added value of *specialised* corpora and *specialised* corpus studies and a small investment in them? Somewhat paradoxically, specialised corpora help all of us to understand how very basic words in English, not just those relating to health and illness, often come to be used in different ways as a result of the different perspectives involved in both specialised and everyday contexts.

In particular, specialised corpora can be a very helpful resource when attempting to master the basic characteristics of scientific discourse. They represent a shortcut to understanding why the grammar rules you learnt at school are so frequently broken when imparting scientific knowledge. One reason for this is that, besides grammatical rules, scientific discourse needs to incorporate and respect power relationships, social, cultural and ethical conventions, sociosemiotic and intercultural factors as well as assumptions about shared knowledge and much more besides. Specialised corpora provide many examples of the interplay between grammar rules and discourse norms and help all of us to understand why spoken and written scientific discourse are so divergent from each other and why mastering their differences is not merely a question of learning specialised lexis. Indeed, as the studies presented here constantly illustrate, understanding how specialised knowledge is represented, and constantly defined and redefined in all types of scientific discourse, not just medical discourse, is far more a question of understanding what lies behind the very different nature of written, spoken and multimodal modes of communication and interaction.

In this respect, we feel we have special duty to encourage students in the later years of their university study, especially Ph.D. students, to invest in specialised corpus studies. Indeed, in different ways and in different circumstances, each of the editors has come to appreciate the value of experimentation that engages students in the construction of specialised corpora, in particular, as recorded in several papers in this volume, in relation to multimedia corpora. Indeed, over the coming years, we can expect specialised corpora to increasingly incorporate viewings of videos, and for university studies in general to deal with far more complex combinations of written, oral and visual discourse than was ever before the case.

Every student should be aware of the digital and multimodal skills that society will increasingly place on them in their careers, a matter that the *Common Framework for Intercultural Digital Literacies* (Sindoni *et al*. 2019)

underscores. In this respect, we feel we have a special duty to encourage all students – whether undergraduate students engaging with the complexities of digital course, for example as part of their translation studies or, for instance, doctoral students engaging in research into specialised discourse in English – to pay special attention to corpus-based multimedia and multimodal studies. Many such studies, as suggested in this volume, and indeed elsewhere, are likely to be undertaken in the field of Medicine and Healthcare, given that the Internet has become a forum where everybody has something to say on these issues.

3. Our message to university colleagues

We have a final duty which is to thank our many university colleagues who have encouraged us to produce this volume and guided us in its execution. There are still, comparatively speaking, few corpus studies that deal with specialised corpora in Medicine and Healthcare, despite the fact that these domains represent a major part of university activity, often including hospital-based care and research. Likewise, there are still few specialised multimodal corpora. As Anne Wichmann observed over ten years ago:

> The technology for recording and storing multimodal data is available, as is software that allows multiple annotations. However, as long as corpus research is driven (and funded) on the basis of just a few aspects of research, such as grammar and lexis, much valuable information could be lost. This paper is therefore a plea to all corpus developers to look beyond their immediate needs and to be a little more visionary in their approach. (Wichmann 2007, p. 86.)

Vision is indeed everything. As this volume goes to press, the third decade in the 21st century is about to be ushered in. We hope this volume, and others that follow it in this and other series, will encourage further investment in specialised corpora in the coming years especially where the relationships between different modes of meaning-making are taken into consideration and where students at all levels, regardless of the degrees in which they are enrolled, are encouraged to take part in specialised corpus-based studies and activities in the furtherance of their digital competences and career prospects.

Lecce, December 6, 2019

References

Sindoni M.G., Adami E., Karatza S., Marenzi I., Moschini I., Petroni S., Rocca M. 2019, *Common framework of reference for intercultural digital literacies*, https://www.eumade4ll.eu/common-framework-of-reference-for-intercultural-digital-literacy/ (07/12/2019).

Wichmann A. 2007 *Corpora and spoken discourse,* in Facchinetti R. (ed.), *Corpus linguistics 25 years on*, Rodopi, Amsterdam, pp. 73-86.

INTRODUCTION TO THE VOLUME

ANTHONY BALDRY

1. Cultural perspectives and starting points in the analysis of medical discourse

This volume brings together five selected papers on medical discourse from the *Clavier 17 – International Conference Representing and Redefining Specialised Knowledge* held in Bari from November 30th to December 2nd 2017. The conference website drew attention to a theme – the capacity of specialised knowledge and discourse to influence our everyday lives – which I wish to examine in this Introduction. In particular, I want to suggest that specialised medical corpora, such as those presented in the papers collected here, provide a framework that helps those engaging with medical discourse to determine how the everyday and the specialised combine to shape the discourse of medical professionals and non-medical communities in relation to both long and short-term factors. Naturally, this includes those cases where the influence runs in the opposite direction where, that is, our everyday lives and needs affect specialised discourse. These opposing trends are one reason why contemporary Medicine is such a vast canvas of expectations, activities and discourses which, if they are to be properly understood and analysed, need to be addressed and summarised holistically.

Accordingly, my starting point is that the papers contribute, in an exemplary way, to illustrating the shifting boundaries in today's society between the two major poles making up the medical discourse cline: healthcare discourse occupies one end, clinical discourse the other. In my view, while the former records the demand for personalised therapies and individual medical services, the latter documents research into society's collective medical needs. Naturally, innovations in both the theory and practice of Medicine have taken place which simultaneously affect both ends of the cline, often causing the cline's endpoints to move further apart and with the further effect that various points along the cline have come to be redefined in recent years. In particular, evidence-based medicine (EBM), in its various forms:

Doi Code: 10.1285/i9788883051531p11

> has made a clear and probable permanent mark on the face of medicine. The introduction of clinical epidemiology into the daily practice of clinicians has offered a systematized, scientific approach to the practice of medicine. (Sur, Dahm 2011, p. 489)

EBM began to emerge in the early 1990s (Sur, Dahm 2011, p. 487; Zimerman 2013) linked, in its original conception, to efforts to remove bias in medical data. The approach has *inter alia* seen the rise of the *systematic review* genre which, by adopting well-defined inclusion and exclusion criteria, collates, re-analyses and re-interprets data acquired in previous studies. Various tools have emerged to make EBM concepts more widely understandable, one of which is the EBM pyramid, further described in Stefania Consonni's article. As an infographic, the pyramid can be readily found with an Internet search as well as in many online videos which use it to illustrate the principle of hierarchical ranking of evidence levels and the correspondence of each level with a specific type of research article. Thus, the pyramid helps explain some basic EBM principles – why, for example, *case reports* and *case series*, which relate respectively to the clinical history of an individual patient and groups of patients with the same condition, are ranked lowest, while *systematic reviews*, with their greater commitment to bias-eliminating criteria, are placed higher up insofar as they are considered more reliable.

However, EBM is an evolving concept, with many knock-on effects, some aspects of which are relevant to the concerns of this volume, as they characterise the evolution that contemporary medical genres are undergoing. *Reporting guidelines* published by different organisations constitute a first adjustment. Besides providing guidance on general issues such as readability, each presents checklists designed to ensure the inclusion of specific data, a way of ensuring standardised structures in the publication of research. Thanks to these guidelines, the expectation is that evidence will be published in a form that facilitates clinical decisions, permits experiments to be replicated and, above all, ensures that evidence can be more easily incorporated into other types of research article, most notably *systematic reviews.* The most well-known reporting guidelines are: CARE for Case Reports; CONSORT for randomised trials; PRISMA for systematic studies; SPIRIT for study protocols; STROBE for observational studies. First port of call for those readers wishing to explore the characteristics and evolution of *Reporting guidelines* as a genre is the EQUATOR network, an acronym standing for Enhancing the QUAlity and Transparency Of health Research, whose mission:

> is to achieve accurate, complete, and transparent reporting of all health research studies to support research reproducibility and usefulness. Our work increases the value of health research and helps to minimise avoidable waste of financial and human investments in health research projects. (http://www.equator-network.org/)

A second adjustment is the reconceptualisation and redesign of the EBM pyramid. As a markedly visual genre entextualising abstract hierarchical concepts, the original pyramid came with the assumption that, as they were prone to a higher level of bias, the genres at the bottom of the pyramid would be less valid than those at the top (Shaneyfelt 2016, p. 121). Doubts about this assumption have led to suggestions that *systematic reviews* should be separated from other types of evidence, and hence other types of research article, on the grounds that they are not the apex of the pyramid but rather a tool for re-analysis and inspection. Hence the publication of reshaped pyramids where *systematic reviews* are represented as a magnifying lens superimposed on a truncated pyramid through which evidence is viewed and re-examined. One result is the renewed pyramid's greater applicability in a wider range of contexts and easier access to the principles it encapsulates:

> This pyramid can be also used as a teaching tool. EBM teachers can compare it to the existing pyramids to explain how certainty in the evidence (also called quality of evidence) is evaluated. It can be used to teach how evidence-based practitioners can appraise and apply systematic reviews in practice, and to demonstrate the evolution in EBM thinking and the modern understanding of certainty in evidence. (Murad *et al*. 2016, p. 127)

A third adjustment stems from the consideration, that while the preference for certain types of research article, for example *systematic reviews* over *case reports*, is not in itself questioned, it is nevertheless subordinate to the principle that:

> Judgment is necessary for interpretation of all evidence, whether that evidence is high or low quality. (Guyatt *et al*. 2008, p. 925)

In this respect, the *Grading of Recommendations, Assessment, Development and Evaluation* system, better known as GRADE, is another corrective, based on the observation that rating the quality of evidence is not the same as grading the strength of recommendations that need to be drawn up and applied in clinical practice. GRADE thus introduces additional forms of ranking and grading:

> To achieve transparency and simplicity, the GRADE system classifies the quality of evidence in one of four levels −high, moderate, low, and very low […] Evidence based on randomised controlled trials begins as high quality evidence, but our confidence in the evidence may be decreased for several reasons, including: Study limitations; Inconsistency of results; Indirectness of evidence; Imprecision; Reporting bias. Although observational studies (for example, cohort and case-control studies) start with a "low quality" rating, grading upwards may be warranted if the magnitude of the treatment effect is very large (such as severe hip osteoarthritis and hip replacement), if there is evidence of a dose-response relation or if all plausible biases would decrease the magnitude of an apparent treatment effect. (Guyatt *et al*. 2008, p. 926)

Understanding the way conflicting principles are resolved in medical discourse might be thought to matter only to those concerned with specialised discourse. In actual fact, the dramatic effects on daily lives arising from reliance on one report or on one principle to the exclusion of others have been a stimulus for rethinking the scientific principles on which EBM is based (Rosner 2012) and how it is applied in healthcare (Wieringa *et al*. 2018). A good example of the care that needs to be taken in the formulation and dissemination of specialised discourse is the debate in online media around HRT (Hormone replacement therapy) where fear is easily aroused:

> Wary of hormone replacement therapy (HRT)? Join the club. Ever since a report by a massive U.S. study called the Women's Health Initiative (WHI) claimed in 2002 that it carried a significant risk of breast cancer and heart disease, most menopausal women remain scared of taking it. Before the alarming news made headlines, around one in four British women was taking HRT. The WHI study's heavily publicised warning sent shockwaves throughout the world. Suddenly a therapy which promised to banish debilitating menopausal symptoms such as night sweats and hot flushes was demonised as a lady-killer. Prescriptions for HRT more than halved in the ensuing two years in the UK, plummeting from around six million a year to just 2.3 million — where the numbers remain today, according to the British Menopause Society. (John Naish, *The Daily Mail*, September 4, 2018)

Ensuring that appropriate safeguards are incorporated in specialised medical discourse as regards the use and definition of words in specialised contexts is, of course, essential, if only because such discourse becomes part of other discourses with significant social consequences. *Randomised controlled trials* (RCTs) are an example:

> There is a danger that the current UK government's interest in RCTs is driven not by their methodological suitability, but because they lend themselves to a model of governance that values context-free quantification and benchmarking. In this situation, RCT advocates would do better by helping build institutions that could put the evidence from trials in its proper context, clarify the conditions under which interventions work or do not work and why, and interpret the meaning of RCTs in relation to plural sources of evidence. This requires engagement across science and politics, alongside an acknowledgement that evidence for policymaking requires expertise as well as data. The new RCT movement needs to grasp this message if it is to benefit the lives of those who are the subject of policy interventions. (Pearce, Raman 2014, p. 398)

Accordingly, various issues relating to interpretation of specialised medical discourse are dealt with in all the papers collected here. The appropriate handling of evidence in scientific discourse is, for example, expertly dealt with in Sabrina Fusari's paper on meat and its carcinogenicity which investigates how medical and other evidence comes to be interpreted and misinterpreted in Public Health. In particular, her paper confirms the significance of the correct interpretation of guidelines and ranking systems when handling medical data. At the same time, this paper, like the others in this volume, is also a demonstration of the significant role that specialised corpora and specialised uses of general corpora play in exploring medical evidence in terms of

contradictions in categorisation, differences in research goals and methods, as well as potential misunderstandings and manipulations.

While *data certainty*, or *data confidence* as it is often referred to in medical publications, has become a distinctive benchmark for the clinical end of the cline, thanks to EBM's quest to validate and certify data quality, the opposite end of the cline has, at the very same time, undergone substantial change, in particular, as regards responses to patient needs and demands. Smart patient-centred technologies, whether concerned with integrating specific devices such as smartphones into healthcare systems (Agarwal *et al*. 2010; Ventola 2014) or with developing AI-based solutions for specific services such as AI-assisted Picture archiving and communication system**s** (PACS) (Alberich-Bayarri 2017), have produced unprecedented healthcare benefits in the management of acute and chronic conditions. In the process, they have given rise to terms such as Precision, Individualised or Personalised Medicine, each of which tend to reflect specialised interpretations of the concept of *individual* that merit further attention in discourse analysis and corpus studies. This is the case, for example, in the field of diabetology (Coons *et al*. 2017; Jameson, Longo 2015; Saucier *et al*. 2017; Swan 2009) where, besides *individual patients*, terms like *patient-centred, precision, individualised* or *personalised* often refer to *individual communities* and what they share as well as what distinguishes them from each other. This is typical of the descriptions of the merger of continuous glucose monitoring (CGM) and insulin pump technologies and the effects this development has on different communities:

> Progressively more accurate and precise, reasonably unobtrusive, small, comfortable, user-friendly devices connect to the Internet to share information and are *sine qua non* for a closed-loop artificial pancreas. CGM can inform, educate, motivate, and alert people with diabetes. CGM is medically indicated for patients with frequent, severe, or nocturnal hypoglycemia, especially in the presence of hypoglycemia unawareness. […] When continuous glucose monitoring (CGM) first became commercially available in the year 2000 its measurement error was more than ±20%. Today, overall measurement error has been reduced by twofold (±10%), and accuracy continues to improve. Size, weight, complexity, and cost of CGM sensors/devices have decreased, whereas the duration of use, specificity, user-friendliness, user interface and displays, data management, and software for data analysis have improved. Numerous studies have demonstrated clinical benefits in multiple patient populations – pediatrics, adolescents, and adults, type 1 and type 2 – with various levels of glycemic control at baseline. Benefit is directly proportional to frequency of use. The effectiveness of CGM can be synergistic with the benefits associated with insulin pumps. (Rodbard 2016, p. S2-3)

One effect of the often slow progress in technological innovation – work on CGM technologies began as long ago as the 1960s (Aathira, Jain 2014) – is, of course, a concomitant desire for technological advances that speed up an end to affected communities' suffering. The advent of wearables, such as fitness trackers and smart bracelets, rings and watches, as well as mobile devices such as smartphones and tablets, represents a tangible expression and part fulfilment of this dream as *inter alia* such devices provide greater mobility and freedom

for both patients and caregivers. In this role, they straddle the subtle boundaries between satisfying consumer needs and healthcare needs:

> you might not realize just how many other things wearable devices can measure. Some of the things smartwatches and activity trackers can measure are downright strange — such as fertility and diabetes — while others are useful to most consumers even though you likely didn't know about them before. (Silbert 2019, p. 1)

The evolution of technology and the changes it brings about are, in my opinion, an under-rated field in corpus studies considering that we live in an age in which mobile technologies have profound effects on everyday lives but which are also linked to mHealth's creation of specialised meanings. This brings with it a strong potential for meaning change in very basic words, most obviously in the relationships between *consumer*, *caregiver* and *patient*. Location tracking devices, which use *GPS, Cell ID and Google Wi-Fi Touch* to track people who need to be protected, are already on the market and include devices for those suffering from AD (Alzheimer's) or other types of dementia (Surendran *et al*. 2018). They respond to a caregiver's *find-you* desire to protect an elderly relative. As such, while some of these devices are clip-on attachments, pendants and wrist bracelets that replicate consumer-oriented wearables, others take the less fashion-conscious form of shoe implants and ankle bracelets, a sign that the subtle boundary between consumer and genuine healthcare needs is being crossed. A tiny, limited step this may be but one that heralds the potential to meet various needs on the constantly expanding healthcare-clinical research cline given that from a clinical research perspective, wearables also represent an opportunity to consolidate the development of predictive digital biomarkers for neurological disorders such as AD, since sensors can be used to record subtle changes over long as well as short timespans. Thus, besides recording slower driving speeds and shorter travel distances, both suggestive of cognitive impairment, sensors can also be used to record changes in gait metrics, sleep patterns, eye movements, pupillary reflexes and disruptions to the brain's cholinergic system, all part of the goal of monitoring many individuals in order to gather evidence of typical patterns that allow more confident diagnoses to be made in the early stages of such diseases:

> In the quest for gold standards for AD assessment, there is a growing interest in the identification of readily accessible digital biomarkers, which harness advances in consumer grade mobile and wearable technologies. (Kourtis *et al*. 2019, p. 1)

Wearables are thus a tangible indication that the healthcare-clinical research cline now stretches from consumer products to clinical research based on mHealth and Big Data (Istepanian *et al*. 2018). The dream of blending social and medical functions within a single device comes, however, with considerable

debate about the complexities entailed – clinical, regulatory, ethical, legal, respect for privacy – that change, condition and constrain the interactions and their interpretations that those concerned, be they doctors, patients, caregivers or other parties, expect to engage in. This is exemplified in evaluations of wearables for neurological disorders (for Alzheimer's, see Ienca *et al*. 2017; for epilepsy and Parkinson's, see Ozanne *et al*. 2018 and Johansson *et al*. 2018; for the role of affective computing in autism, epilepsy, and sleep memory formation, see Picard 2014). The rise of both wearables and IoT (Internet of Things) in healthcare (Metcalf *et al*. 2016; Yuehong *et al*. 2016) suggests that the pace of change is accelerating and will continue to do so, creating new expectations about the fulfilment of healthcare dreams in the management, and above all, self-management of chronic illness. This is the case for instance with insulin-dependent diabetics, where there is a particularly strong awareness that the day when the closed-loop artificial pancreas will provide flawless non-stop automated coverage is drawing closer and closer (Breton *et al*. 2012; Clarke *et al*. 2009):

> Thanks to the effective integration of engineering and medicine, the dream of automated glucose regulation is nearing reality. (Doyle *et al*. 2014, p. 1191)

While better self-management and innovations in chronic care delivery systems (Chiauzzi *et al*. 2015; Milani, Lavie 2015) have contributed to reducing the stresses associated with chronic illness, the medical dreams in question transcend healthcare self-management and affect all aspects of the healthcare-clinical research cline. Diabetology is again a good example. In the process of transcending the capabilities of the natural pancreas, artificial pancreas technology has encouraged other dreams and the process of their fulfilment. Combined CGM and insulin pump technology now hooks up with the smartphone (Lanzola *et al*. 2016), meaning that data can be sent directly to remote patient monitoring systems in hospitals which, in their turn, feed the data pool that allows clinical research to achieve even higher standards of data confidence. This process makes use of, and strengthens, Remote Monitoring [RM], less prominent today in the public eye, but a technology destined, as suggested above, to have an ever greater social and medical impact:

> Although rare at this moment, incentives to use RM technology are likely to increase in the near future as the body of evidence of clinical and/or economic benefit grows. (Rojahn *et al*. 2016, pp. 1-2)

All these examples underpin the impact of technology on both the healthcare and clinical research poles of contemporary Medicine. The fulfilment of medical dreams triggers changes in the way people, in their professional and lay roles, talk and write about medical events, one reason why we need specialised medical corpora that explore, for example, the discourse aspects of

digital healthcare communication in different communities (Hunt, Harvey 2015; Crawford *et al*. 2014). Such studies can give us a better understanding, for example, not just of the role of technological innovations in the realisation of medical dreams, but also of the effects that they have on the frequency and meaning of some very basic words used in Medicine, a matter not underestimated in the papers in this volume, but which, nevertheless, is very much in need of further investigation and consolidation.

What, for example, in contemporary Medicine is the meaning and frequency across different medical specialties of the term *patient*? How frequently in everyday and specialised discourse does it refer not so much to *real* patients, whether viewed individually or collectively, but instead to *hypothetical* ones? The ranks of these imaginary patients have certainly increased as a result of *virtual patient* genres such as *interactive virtual patient scenarios* used in healthcare training (Shah *et al*. 2012) and *simulated hospital patient flows* used to promote cost-effective healthcare management (Heinrichs *et al*. 2008; White 2005). As simulations, they have one foot in the services provided by real hospitals, the other in the world of *what-if* hypotheses and predictions (Trickett, Trafton 2007; Bewley, O'Neil 2013; Reese *et al*. 2010), so that the meanings associated with well-known patient categories such as *hospitalised patients*, *discharged patients* and *recurrent patients* are now dependent on the way these genres, with their inherent ambiguity, are interpreted by different communities, all of which brings us back to the basic question: what effects do digital worlds and digital technologies have on the meaning of basic medical terms?

To what degree, for instance, when used in healthcare simulation services, do words like *simulated* and *standardised patients*, still retain their traditional association with *real* people trained to act out acute or chronic medical conditions in face-to-face contacts with medical trainees (Churchouse, McCafferty 2012)? To put the matter another way, to what extent are the terms *virtual patient* and *simulated patient* now conflated in medical training simulations? How are these terms used in highly specialised contexts where trainees' interactional and clinical competences (Battles *et al*. 2004) are measured, for example, with reference to simulations of various patient categories including, for example, *difficult patient simulations* (Gorini *et al*. 2008; McGrath *et al*. 2018; Rizzo, Talbot 2016; Levine *et al*. 2016)? To judge from a survey of 536 articles published between 1991 and December 2013 of the use of *virtual patient* in healthcare education, alas not carried out within corpus linguistics, or indeed any field of linguistics, such questions *do* get posed but are only partially answered:

> There are potential limitations to our study. The aim of our research was to classify the body of literature about virtual patients. Therefore we focused exclusively on the search term "virtual patient", not including other potentially related search terms, such as "patient simulation". (Kononowicz *et al*. 2015, p. 17)

What stands out in this and many other non-linguistic studies, such as the *systematic reviews* published in medical journals, is either the absence of the word *corpus* or its use in a way that is hard for discourse analysts and corpus linguists to swallow as the term is used merely to describe a set of publications on the same theme in which frequency counts rarely go beyond counting the number of times a specific item recurs, for example, the number of patients within a cohort who can be attributed to specific subgroups. Discourse analysts and corpus linguists will inevitably share the conviction, expressed in various ways in this volume, that the tools and concepts of corpus studies are beneficial in many medical domains, where there is a need to understand the *relations* existing between terms, in particular, their typical distributions relative to each other, a matter successfully explored, for example, in the paper by Stefania Maci and her co-authors. Such studies have significant applications in medical training but alas the message that specialised corpora, their construction and use need to be part of basic medical training in digital and multisemiotic literacy is hard to get across (Baldry 2011) and, alas, even contested. This issue is further discussed, with reference to the papers by Anna Loiacono and Francesca Tursi and by Davide Taibi, Ivana Marenzi and Qazi Asim Ijaz Ahmad, in *Section 3* of this Introduction in relation to corpora as part of simulation services.

The need for more corpus-based studies concerned with basic medical terms, a task that this volume successfully undertakes, is all the more important given that what appear to be everyday words will in fact take on specialised meanings that are frequently the source of misinterpretation and misunderstandings:

> […] the term "virtual patient" is used to describe a multitude of technologies and approaches, making effective communication difficult when educators, researchers and IT specialists share their experiences with VPs. (Kononowicz *et al*. 2015, p. 12)

Somewhat ironically, the expectations that accompany scientific certainty and precision, whether in the field of medical analysis or discourse analysis, seldom avoid the need to reckon with, and measure up to, human nature with all its failings, in particular its tendency to reject and decry the expertise of others when reacting to bad news or sudden illness. Doctors and patients still play the age-old cat-and-mouse game of not trusting each other, of complaining about each other's incompetence and asserting that they know best, a game that has characterised Medicine throughout its history. However, the discourse that surrounds the mutual accusations of fallibility has changed, as is highlighted

in the following scene from an episode in the *House M.D.* TV series, appropriately named *Epic Fail*, where the untrusting patient, Vince, mentions to his doctors, Thirteen and Taub, that he might be suffering from mercury poisoning as he eats a lot of sushi, while they suspect CRPS (complex regional pain syndrome) − incorrectly as the correct diagnosis eventually turns out to be Fabry disease:

> **VINCE**: I don't buy it.
> **TAUB**: CRPS isn't that well understood, but –
> **VINCE**: I think it's mercury poisoning. I eat a ton of sushi.
> **THIRTEEN**: And you're currently getting mixed reviews in "Speed-the-Plow" on Broadway. (Vince and Taub look at her questioningly) Google it. It's pretty hard to consume enough fish to give yourself mercury poisoning, and it doesn't usually present solely with pain.
> **VINCE**: But it can. Check out the "Atlantic Medical Journal". This guy came in with burning pain caused by, uh, "erythromelalgia", caused by mercury poisoning.
> **THIRTEEN**: Who needs actual doctors when you got the internet?
> **VINCE**: No offense, but doctors make mistakes. Medical errors are up 30% this year.
> **TAUB**: You should check the rate of patient error.
> **VINCE**: There's a ton of information out there. Why wouldn't I educate myself, be my own advocate? CRPS came up in my search too. But I've never had any skin discoloration, and my pain is sporadic and not constant. It's got to be worth one lousy blood test.

How different this discourse is from medical interviews from the pre-Internet era. Note, in particular, Vince's references to online sources, including medical journals, his use of acronyms, technical terms and, of course, statistics. These features are consistent with his attempt to undermine the authority and power that derive from specialised discourse, by emulating and, as it were, 'highjacking' it. As such, the scene captures and characterises a typical flashpoint in contemporary Medicine arising from the changed nature of doctor-patient discourse. The discourse has changed because the patients have changed as a result of easy access to specialised discourse that new technologies have made possible. Besides benefits, this comes with a greater potential for loss of trust. This focus is reflected in the episode title *Epic Fail*, often used to describe unexpected and humiliating defeats associated with digital genres, but in Vince's case attributable to the failures in providing a correct diagnosis as well as to the flaws in the video games that he designs that are in fact caused by his illness. Vince's statement is, of course, an extreme form of *do-it-yourself* medicine that characterises this episode's exploration of lay *vs.* technical and specialised discourse. That the changing boundaries between these types of discourse can undermine mutual respect is, of course, well-known to experienced doctors in the real world and not just the TV world. Thus, for instance, one doctor has noted that patients currently undergoing total joint arthroplasty are different from those in the past. Not satisfied with increased wealth, life activity expectation, and life expectancy, they expect miracles – a result of the revolutionary explosion of, access to, and dissemination of information:

> Our patients are citizens of our modern age. Our public has come to expect miracles in medicine as the norm, yet these miracles are not without inherent risk. The trap implicit in allowing an incompletely informed populace to drive the decisions we make may be bridged by a more complete understanding of who our patients are and what their needs include. This discussion attempts to offer some insight into the forces at play. It focuses on how the changes in society, population, and technology have affected patients' knowledge and attitude toward medicine and what our response as physicians should be. (Mason 2008, p. 1)

Naturally, the desire to identify with medical dreams and miracles is part of human nature so much so that it comes as no surprise that in today's digital world the layman is urged to be properly informed about "best practices" in medical treatments and encouraged, for better or for worse, to prise open the sealed box of clinical knowledge and get right inside in order to discover its secrets. Indeed, it is not by chance that Vince mentions being his *own* advocate. While his despair and protests are understandable, his use of this word is a reference to the *patient advocate*, an emergent healthcare professional whose role in circumventing incomprehension and mistrust is defined by the PACB, the Patient Advocate Certification Board, as follows:

> A patient advocate is a professional who provides services to patients and those supporting them who are navigating the complex healthcare continuum. Advocates work directly with clients (or with their legal representatives) to ensure they have a voice in their care and information to promote informed decision making. Advocates may work independently or in medical or other organizational settings. They serve individuals, communities, disease-specific populations, and family caregivers. Synonyms may include health advocate, healthcare advocate, healthcare advocacy consultant, healthcare consumer advocate, and other phrases that imply this role. (*https://pacboard.org/decisions-and-documents/*)

Regardless of doubts about the *patient advocate*'s status, training, true value and future evolution (Schwartz 2002), the rise of such intermediaries shows that compared with the past (Kaba, Sooriakumaran 2007; Conti, Gensini 2008; Harrison 2018), the doctor's addressee is no longer solely the *individual* patient:

> A core challenge of 20th-century medical education was reconciling the clinical care of patients with a scientific approach to medicine. Educators using proposals as diverse as the Flexner Report and patient-centered medicine struggled to ensure the continuous progress and clinical application of medical science while upholding and advancing the ideals, ethics, and art of bedside practice. In 2011, this struggle continues but must give some ground to another challenge: With expanding health care costs and inequities at critical mass, the next generation of physicians must be taught how to integrate population consciousness into clinical practice. […] One might say that we can no longer ignore the other 300,000,000 patients in the room. (Kontos *et al*. 2011, p. 1341)

Studying the structure of medical interviews (Silverman 1987) often from the standpoint of power relationships and asymmetries (Pizzini 1990; Steele *et al*. 1990; Menz, Al-Roubaie 2008) and more generally the doctor-patient relationship in medical discourse analysis (Gotti, Salager-Meyer 2006; Ferguson 2001; Heritage, Maynard 2006) has, of course, been an area of considerable

success for discourse analysis and corpus studies. However, for those studying contemporary medical discourse there is a need to recognise the changed circumstances of 21st century Medicine and to make adjustments. As the contributions in this volume demonstrate, one effect of the stretching out of the two ends of the medical cline is to redefine the doctor-patient relationship, reshaping its empowerments and dualities. Like it or not, the patient is often viewed – and not just in the United States – as a consumer of healthcare services, so much so that, as Fusari's paper demonstrates, much contemporary healthcare documentation is not about doctors and patients but about consumers. The latter are *potential* patients who become *real* patients only when the protective web, spun in many Public Health contexts by national and international institutions such as health and food safety agencies, fails.

Indeed – despite all contemporary Medicine's efforts to shield specific patient communities – through better triage systems (Parenti *et al*. 2014), better patient safety in hospitals (Pronovost, Vohr 2010), better discharge and follow-up procedures and protocols (Naylor *et al*.1999; Gonçalves-Bradley *et al*. 2016; Shoeb *et al*. 2012), or even investment in transitional care for those with continuous complex care needs (Coleman 2003) – the pressure is such that the rope holding together the various points of the healthcare-clinical research cline inevitably snaps. Consumer protection and whistleblowing then step in. Public outcries expose flaws in healthcare services that go well beyond Vince's private face-off with his doctors. Hospital interpreting services based on video links are a classic example of what can go wrong in doctor-patient interaction when speedy access to digital services of the required quality is not available:

> Many deaf patients have taken to social media to complain about the use of video interpreting services in emergency rooms. (https://www.statnews.com/2017/05/22/deaf-patients-interpreters/)

Indeed, such protests are the first step in a process that leads to the courts upholding the right to interact as an integral part of healthcare services:

> regardless of whether a patient ultimately receives the correct diagnosis or medically acceptable treatment, that patient has been denied the equal opportunity to participate in healthcare services whenever he or she cannot communicate medically relevant information effectively with medical staff. (US Court of Appeals, Case: 16-10094, p.14, 05/08/2017; http://media.ca11.uscourts.gov/opinions/pub/files/201610094.pdf, page14)

The right to interact is, of course, conditioned by the way healthcare services are structured. When video interpreting services fail because doctors and hospital healthcare workers do not know how to operate the equipment or when poor screen quality effaces the meaning-making resources on which sign language depends, the spotlight inevitably falls on the ties between interaction, service planning and, above all, teamwork (Keating, Mirus 2003). As a further

consequence, it also falls, more indirectly, on the need to redefine the nature of interaction in medical contexts in a digital society.

However, the right to interact also affects doctors and other healthcare workers just as much as it affects patients and their caregivers. Thus, one answer to the issue of not knowing what to say to patients who expect miracles lies in teamwork and, of course, training to become part of a medical team, something that has often been shown to be a significant response when it comes to rare diseases. CRPS, complex regional pain syndrome, *is* indeed, just as the scene in *Epic Fail* portrays it, a rare disease that is hard to diagnose and hard to treat as it is a poorly understood condition causing persistent, severe and debilitating pain. The 21st century has seen the rise of multidisciplinary approaches that exemplify efforts to bring the healthcare and clinical research aspects of contemporary Medicine closer together; in the case of CRPS, this has been done by strengthening, on the healthcare side, patient education, self-management, physical rehabilitation, pain relief and psychological support and, on the clinical research side, by promoting new configurations of specialist knowledge:

> We have learned much about CRPS in the past 10 years, and we have been given a glimpse into some treatments that for the first time, promise effective pain reduction for those with long-standing disease. The quality of clinical trials has much improved and the quantity of research into this condition has skyrocketed. While we still do not know what causes CRPS, one has the sense that efforts to tackle this fascinating, debilitating condition are exemplary for the progress of the new field of Pain Medicine to come into its own. (Goebel 2011, p. 1747)

Teamwork combined with innovation in both techniques and technologies is a game-changing aspect of contemporary Medicine that leads to new medical specialties that successfully link up both ends of the medical cline. This is a recurrent feature of contemporary Medicine that needs to be fully grasped when contemplating the construction and use of specialised medical corpora. We have already outlined the effects of Personalised care in the field of diabetology, but can further exemplify the interplay between changes in medical service culture and teamwork in relation to drug therapy and its use of automated delivery systems (Goundrey-Smith 2019). This field is characterised by transitions to new areas of clinical research:

> Individuals respond differently to drugs and sometimes the effects are unpredictable. Differences in DNA that alter the expression or function of proteins that are targeted by drugs can contribute significantly to variation in the responses of individuals. Many of the genes examined in early studies were linked to highly penetrant, single-gene traits, but future advances hinge on the more difficult challenge of elucidating multi-gene determinants of drug response. This intersection of genomics and medicine has the potential to yield a new set of molecular diagnostic tools that can be used to individualize and optimize drug therapy. (Evans, Relling 2004, p. 464)

However, it also affects the other end of the cline thanks to the introduction of the electronic prescription (EP) of drugs, all part of the era of paperless hospital healthcare systems:

> NHS hospitals in England are expected to be paperless by 2020 as set out in a comprehensive framework published by the National Information Board. The use of hospital electronic prescribing (EP) systems is therefore likely to increase rapidly in the near future. The aim of this review is to summarise the available evidence of the impact of inpatient EP on patient safety, with a focus on implications for the UK. […] The review concludes with considerations of the evolution of EP in healthcare, especially in relation to advances in health information technology, inpatient involvement with their medication in the context of EP, and how EP may be used by policymakers and end users to further benefit patient safety. (Ahmed *et al*. 2016, p. 1758)

The rapid increase in the use of integrated EP and RD (robotic dispensing) in hospitals (Beard 2017; Crawford *et al*. 1998) shows that while some aspects of the IOT hospital are already part of the here and now, many others, despite many challenges, are imminent (Laplante 2016) including changes in the way interactions between people are envisaged. Indeed, Medicine is clearly in a state of transition towards complex forms of teamwork that tie together all aspects of the medical cline in a way that ensures all forms of discontinuity are avoided. Capturing this transitional state through corpus studies, which, includes, of course, the construction and design of multimedia corpora such as the one described in the paper by Davide Taibi, Ivana Marenzi and Qazi Asim Ijaz Ahmad, will provide a better understanding of the influence that specialised knowledge and discourse have on our everyday lives.

Indeed, when I re-read the papers in this volume, I really feel that these corpus-based studies are helping to pinpoint cases where doctors and clinicians are caught between opposing demands, such as those generated by the varying interpretations of EBM that I have described above, which go a long way to defining contemporary Medicine. As such, I feel that the papers rightly go beyond many traditional studies of medical discourse with their focus on *direct* forms of interaction – such as the analysis of doctor-patient medical interviews (Schegloff 1999; Ong *et al*. 1995; Maynard, Heritage 2005) or the analysis of the structure of research articles written for and read by medical elites (Hopkins, Dudley-Evans 1988; Swales 1990; Salager-Meyer 1991; Hyland 1998). There is an urgent need to map out the more mediated and indirect forms that characterise today's medical discourse and to focus on teamwork in contemporary Medicine, highlighting the involvement of non-medical professionals whose contribution is nevertheless fundamental to the promotion of healthcare services and clinical research. All this requires discourse and corpus studies to be related to the cultural, philosophical, organisational and technological aspects of contemporary Medicine as well as the purely textual. Only when this wider perspective on medical discourse is embraced, will it be

possible to really understand all aspects of how specialised medical knowledge and discourse are influencing our everyday lives.

2. Using specialised corpora to explore transition and teamwork: terminological, textual and interactional aspects

If I have not yet presented the papers collected in this volume, it is because I wanted to set the stage with my vision of contemporary medical discourse as a set of transitions and negotiations between convergent and non-convergent interests that cover an ever-expanding constellation of contexts. Although no reference is explicitly made in the papers published in this volume to the healthcare *vs.* clinical research cline that I have characterised above, I find it hard *not* to interpret them collectively as different perspectives on this cline. In their various ways, the papers deal with a range of professional figures who participate in discourse communities that occupy various points along this cline, each with their own discourse styles and each needing to make adjustments when engaging with other communities, given the multiplicity of audiences and addressees that contemporary medical discourse needs to take on board.

As they browse through this volume, readers, especially those exploring corpus-based approaches to medical genres for the first time, will appreciate the value of having five very different illustrations of specialised medical corpora in a single volume. The intriguingly dissimilar choices the authors have made as regards the type of corpora they have used and the type of linguistic and textual units they have chosen to explore, chime with my belief that careful reflection on starting points in the analysis of medical discourse is essential given the very varied cultural frameworks in which contemporary Medicine works. Thus, while Stefania Consonni's paper – short title *HIV Discourse in the British Medical Journal, 1985-2005* – is based on a corpus of research article *titles* that appeared in the *BMJ* in relation to HIV over a 20-year period, Sabrina Fusari's paper – *Does Meat Cause Cancer?* – instead uses a corpus assembled from a range of academic journals featured in the database *Elsevier Science Direct* to explore the relationships between cancer and food in terms of *collocations* and *collocational patternings.* While both these papers are based on small, highly specialised corpora created by the authors, the paper by Stefania Maci, Réka Jablonkai, Marek Łukasik, Sophiko Daraselia and Daniel Knuchel – *Disambiguating Near Synonyms in Medical Discourse* – uses the *BNC* to examine the distributional characteristics of *three lemmas,* specifically *illness*, *disease* and *sickness* – in terms not just of the differences arising from their use as singular/plural *lexical items* but also, in terms of the

semantic profile emerging from these terms' deployment in the different contexts implicit in contemporary Medicine, which, of course, includes lay/professional contrasts.

The paper by Anna Loiacono and Francesca Tursi – *Mapping Medical Acronyms* – focuses on the trials and tribulations facing medical students when learning, and learning about, medical *acronyms,* and provides an important snapshot of students transiting from lay discourse to professional discourse. Learning to cross that Rubicon also requires an ability to look back and reflect critically on the effects of selecting one discourse style over another and to understand that, while abbreviatory strategies (acronyms included) are typical of professional discourse, as well as being both culture-dependent and language-dependent, they are nevertheless increasingly being mastered by patients, caregivers, consumers and other non-medical professionals who incorporate them into their everyday 'semi-professional' discourse. The acronyms described in this paper were extracted from the *House Corpus,* whose derivation from the well-known TV series is described in the final paper by Davide Taibi, Ivana Marenzi and Qazi Asim Ijaz Ahmad. As suggested by the title – *Ain't that sweet: Reflections on scene level indexing and annotation functionalities in the House Corpus Project* – this paper explores the construction of a multimedia corpus around a further but somewhat unusual unit of analysis in corpus studies: *scenes.* As such, the paper considers the value of the scene as a meaning-making unit, when using a specialised corpus as a form of simulation, in other words, as a way of exploring simulated activities in medical and language-related training activities in universities. In so doing, it lays the bases for exploring the still uncharted waters of the relationship between corpora and the world of simulated medical services that I have referred to above and further describe in *Section 3* of this Introduction.

No two papers in this volume consider medical discourse in the same way. So just where do these papers fit on the healthcare-clinical cline sketched out above? Although the papers are published in alphabetical order based on the initial letter of the first author's surname, other more meaningful distributions suggest themselves, for example, the issue of the contribution that the papers make to language variation in corpus studies. They do so in a way that does not question the centrality in medical discourse of the medical interview or the research article but which, nevertheless, implies that other forms of medical discourse, in particular discourse that is spoken, written-to-be-spoken and written-for-non-specialists, need to be investigated. This a first step in ensuring that genres such as the medical interview and the research article are studied in terms of the way they meet up with and interact with other forms of medical discourse (Morris, Chenail 2013; MacDonald 2002; Zabielska 2015). In this respect, transition is a keyword when analysing

contemporary medical discourse as it affects so many of the basic terms whose meanings are too often assumed as being in some way fixed.

While this is not the place to provide a full semantic history of the words *clinic*, *clinical* and *clinician*, their changes in meaning amply illustrate how a change in cultural perspective can cause meanings to shift from one end of the medical cline to the other. In the 17th century, *clinic* meant a "bedridden person, one confined to his bed by sickness," (source: www.etymonline.com/word/clinic), a *patient-centric* standpoint, which helps us to understand and appreciate both the mid-19th century the use of the term *clinician* as "one who makes a practical study of disease or sick persons," (source*:* https://www.etymonline.com/search?q=clinician) and, in addition, the subsequent extension to the teaching of medical students that we find in the online OED's definition "Of or pertaining to the sick-bed, *spec.* to that of indoor hospital patients: used in connection with the practical instruction given to medical students at the sick-beds in hospitals". This meaning is partly the result of the work of William Osler (1849-1919), the first to bring medical students out of the lecture hall for bedside clinical training:

> The medical clinic instructional model that Osler put into effect revolutionized medical teaching in the art and science of diagnosis and patient care. […] Medical students became actual members of the patient care team, taking histories, doing physicals, doing the laboratory work, and making rounds with the residents and faculty. Thus evolved the medical clerkship, which was extended to surgery, obstetrics and gynecology. This clerkship did for the clinical students what laboratory work did for the scientists. (Walker 1990, p. 19)

While the above quotation – and in addition other sources such as the entry for *clinic and clinical* in the online OED – show that towards the end of the 19th century *clinical* referred to observations made about individual patients, often in a teaching context, and a *hospital clinic* was the place where this was carried out, today in many English-speaking countries a *hospital clinic* increasingly refers mostly to a medical centre for *outpatients*, unlike other cultures and languages which use cognate forms of this word to refer to the wards in which hospitalised patients are looked after or, more abstractly, to the science of treating such patients. While medical training still continues in such outpatient clinics, a further break with the past is that explicit patient consent is often required as regards medical trainees' right to be present. As the State of Victoria's website for *Specialist clinics in hospitals* demonstrates, patients' rights prevail and are indeed supported, where necessary, by the intermediation of *patient representatives*, a further indication that modern healthcare is much more than just the doctor-patient relationship:

> These specialist clinics, which are sometimes referred to as 'outpatients', are for people who are not currently admitted to the hospital. […] Patients may be seen by a range of health care professionals, including students from allied health, nursing and medicine who are in different stages of their training. Public hospitals are teaching hospitals and it is intended that students

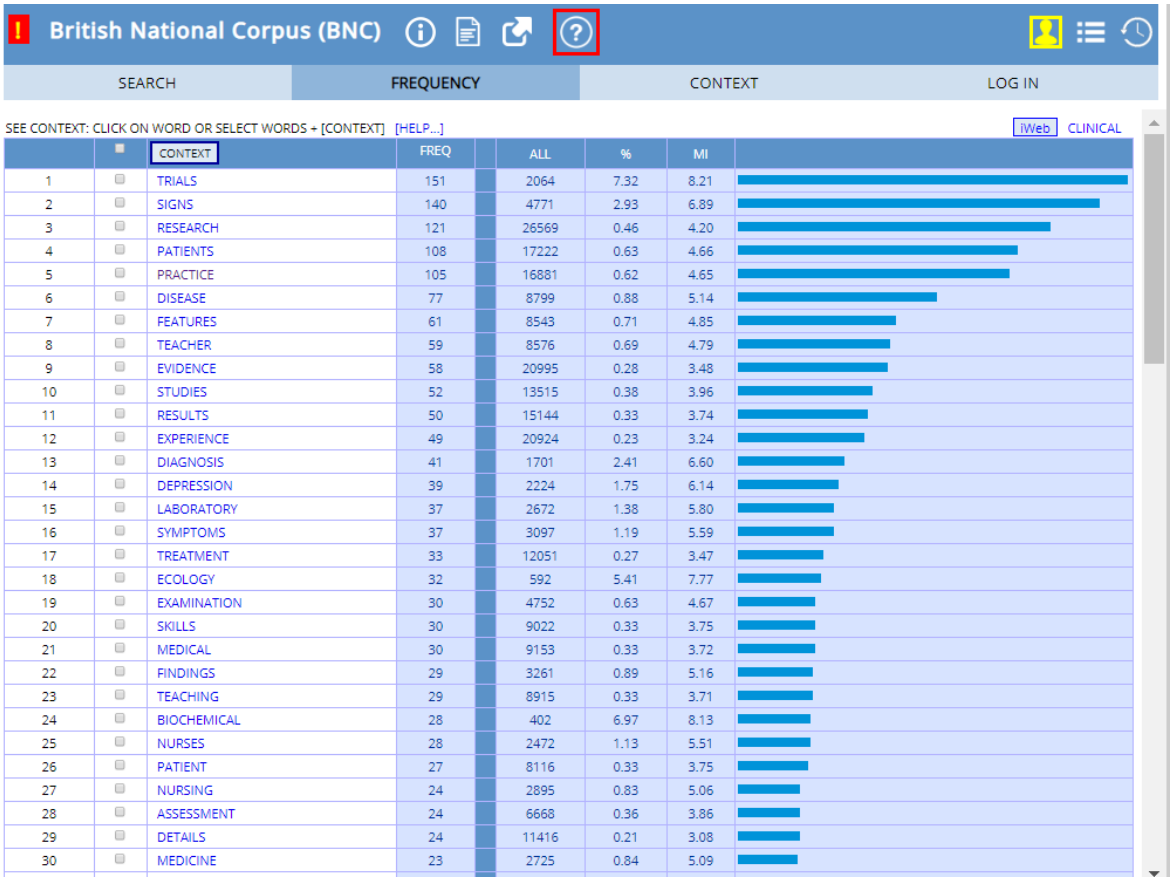
> interact with patients to increase their clinical knowledge. However it is your right to refuse to be seen by a student. Your doctor should introduce these staff to you. If you do not want additional staff present please let the doctor know. This will not affect your care in any way. […] Every Victorian public hospital has a patient representative, and their name and telephone number can be provided to you by the health service. […] The patient representative will work with you to find a resolution to any complaint, or, if necessary, investigate the matter further. (www.betterhealth.vic.gov.au/health/servicesandsupport/specialist-clinics-in-hospitals)

To my mind, a major role of specialised corpora, underscored many times in this volume, is to point out transitions to new meanings and to warn that, while today's dictionary entries are themselves evidence-based and illustrated with examples taken from carefully-constructed corpora, this does not necessarily mean that the definitions they provide have been updated to reflect contemporary meanings. In this respect, dictionary definitions of *clinical* are rather interesting. For example, the online Merriam-Webster's definition of clinical –"*of, relating to, or conducted in or as if in a clinic such as* a*: involving direct observation of the patient: clinical diagnosis*; b: *based on or characterized by observable and diagnosable symptoms: clinical treatment, clinical tuberculosis see also clinical psychologist*"–has certainly moved on from the online OED's "*Of or pertaining to the sick-bed"* definition quoted more fully above; as the OED is a historical dictionary, this is perfectly in keeping with the *end-of-the-19th-century* definitions already mentioned. Note, however, that the Merriam-Webster definition still focuses on the *observable* and the *direct*, the latter also the major focus in the online Collins dictionary: "*Clinical means involving or relating to the direct medical treatment or testing of patients".*[1]

While the words *based on* at the start of the second part of Merriam-Webster's two-part definition *do* point to this word's extension into the research world's analysis and aggregation of data, the concept of *direct* observation, treatment or testing, which both Webster and Collins underscore, would appear to be at odds with many of the examples given. Thus of the 32 examples quoted in Collins, mostly taken from *The Times*, *The Sunday Times* and *The Sun*, the most frequent collocation is *clinical trials* (12/32). Quite apart from suggesting that the expression *clinical trials* is now part of the everyday knowledge of these newspapers' lay audience, the very fact that in more than a third of the examples the observation in question relates to *data* and not to *patients* suggests that the everyday meaning today is a far cry from bedside collocations such as *Clinical Professor; clinical medicine, clinical clerk, clinical instruction; clinical thermometer; Clinical Clerkship* that we find in the OED's 19th century examples. Indeed, TV medical soaps apart, *bedside*

[1] OED: https://www.oed.com/view/Entry/34381?redirectedFrom=clinical#eid; Merriam-Webster*:* https://www.merriam-webster.com/dictionary/clinical; https://www.collinsdictionary.com/dictionary/english/clinical.Retrieved 28.06.2019

clinical lectures appear to be giving way to less theatrical and more mediated forms of doctor-patient interaction.



	CONTEXT	FREQ	ALL	%	MI
1	TRIALS	151	2064	7.32	8.21
2	SIGNS	140	4771	2.93	6.89
3	RESEARCH	121	26569	0.46	4.20
4	PATIENTS	108	17222	0.63	4.66
5	PRACTICE	105	16881	0.62	4.65
6	DISEASE	77	8799	0.88	5.14
7	FEATURES	61	8543	0.71	4.85
8	TEACHER	59	8576	0.69	4.79
9	EVIDENCE	58	20995	0.28	3.48
10	STUDIES	52	13515	0.38	3.96
11	RESULTS	50	15144	0.33	3.74
12	EXPERIENCE	49	20924	0.23	3.24
13	DIAGNOSIS	41	1701	2.41	6.60
14	DEPRESSION	39	2224	1.75	6.14
15	LABORATORY	37	2672	1.38	5.80
16	SYMPTOMS	37	3097	1.19	5.59
17	TREATMENT	33	12051	0.27	3.47
18	ECOLOGY	32	592	5.41	7.77
19	EXAMINATION	30	4752	0.63	4.67
20	SKILLS	30	9022	0.33	3.75
21	MEDICAL	30	9153	0.33	3.72
22	FINDINGS	29	3261	0.89	5.16
23	TEACHING	29	8915	0.33	3.71
24	BIOCHEMICAL	28	402	6.97	8.13
25	NURSES	28	2472	1.13	5.51
26	PATIENT	27	8116	0.33	3.75
27	NURSING	24	2895	0.83	5.06
28	ASSESSMENT	24	6668	0.36	3.86
29	DETAILS	24	11416	0.21	3.08
30	MEDICINE	23	2725	0.84	5.09

Figure 1
Context of Use Map for *clinical*; British National Corpus: www.english-corpora.org/bnc/.

Support for the belief that dictionary definitions of *clinical* ought to recognise that the term *clinical* no longer pertains exclusively to the healthcare end of the medical cline, but now also embraces quite extensively clinical research at the opposite end with its far more indirect and abstract relationships, comes from the *British National Corpus (BNC)* where, as shown in Figure 1, the high frequency of use of *clinical* in relation to *trials, signs, research* can be compared with much lower rankings for direct observations and interactions with patients suggested by collocates such as *nurses, nursing, assessment, details.* The contrastive ranking of *patients* (*i.e.* collective) and *patient* (*i.e.* a specific patient) is particularly noteworthy.

The closer inspection that specialised corpora provide brings further important confirmatory evidence, specifically from the *House Corpus* described in two of the volume's articles. The *House M.D.* TV series is a modern-day reconstruction of *clinical* in the Victorian sense of a clinician as a bedside sleuth epitomised by the Edinburgh-trained physician Conan Doyle,

whose most famous literary creation, Sherlock Holmes, is unquestionably the basis for Greg House, the lead clinician in this TV series (Mamatas 2007). Given that the series focuses on House's brilliant diagnoses of the rare conditions that his patients are suffering from, as well as his 'lecturing' of his medical team who have missed vital diagnostic clues, the *House Corpus* could have been expected to show that this TV series makes a very large use of the word *clinical*. In fact, it shows the opposite. The term occurs in just 25 of the 6300 or so scenes, with an overall score of 29 tokens, 21 of them in the expression *clinical trial(s)*. By contrast, the term *medical* appears in 323 scenes (395 tokens) and *clinic* in 245 scenes (348 tokens).

Naturally, it can be objected that a TV medical soap is not a true reflection of today's everyday or specialised discourse. However, other sources and considerations support the view that the rise of EBM has left an 'indelible mark' on the term *clinical* causing its meaning to change. EBM's scientific and cultural role in this process of semantic change can perhaps be best appreciated when viewed, in a diachronic perspective, as the most recent stage in the much longer textual and terminological journey that *clinical trials* (and the methods used to acquire and report data) have undergone – which brings us to another major port of call: *The James Lind Library: www.jameslindlibrary.org/*:

> To illustrate the evolution of ideas related to fair tests of treatments from 2000 BC to the present, the James Lind Library contains key passages and images from manuscripts, books and journal articles, many of them accompanied by commentaries, biographies, portraits and other relevant documents and images, including audio and video files. New material is being added to the website continuously, as relevant new records are identified and as methods for testing treatments evolve. (Chalmers *et al*. 2008, p. 259)

Regardless of whether we consider this site as a corpus or 'merely' a fascinating collection of texts relating to the rise, evolution and fortunes of *clinical trials*, the site's timeline search tools are sufficient to allow counts to be made in the *Records* section for the presence of *clinical* in text titles. There are none prior to the 18th century, 2 out of 22 records in the 18th century; 4 out of 24 in the 19th century, 23 out of 93 in the first part of the 20th century and then a massive increase to 75 out of 157 in the second part of the century. The decline in the 21st century – only 4 out of 29 – is partly due to the fact that the word *clinical* is omitted as the term *trials* is considered distinctive in itself and partly due to the fact that subtler classifications are now incorporated into research article (RA) titles. As Consonni puts it in her article: "compound titles allow readers and fellow researchers to rank the evidence provided in the RA within the EBM hierarchy" and thus determine "what impact its results can be expected to have in terms of methodological credibility".

Without wishing to labour the point any further, there is a need for specialist corpora to examine the meaning of basic terms such as *clinical* in medical websites that explain *clinical trials* to laymen. In this respect, we

should recall that EBM divides *clinical studies* into *observational studies* and *clinical trials.* While the distinction may be clear to medical professionals, considerable effort is needed to explain their distinctive functions to lay persons, in particular when promoting participation in *clinical trials* which, unlike *observational studies,* crucially depend on recruiting volunteers *not* participating in any other clinical trials. Besides explaining eligibility criteria, dedicated websites thus undertake the task of explaining the rules of the game but also coax lay persons into overcoming their reluctance to participate in clinical trials with reference to the benefits for others with the same social and/or medical status, which is why we find the US *National institute on Aging*: (www.nia.nih.gov/health/what-are-clinical-trials-and-studies) giving an age-related example, while the *Cystic Fibrosis Foundation* (www.cff.org/) presents a disorder-related example.

While slogans such as *Help us blaze a trail to better treatments and a cure for CF* are indicative of the promotional techniques used in advertising and marketing discourse in today's highly specialised medical interpretations of social advertising, such persuasion differs markedly from that traditionally associated with medical 'healthscare' campaigns (Baldry 2005, pp. 45-63; Baldry, Kantz 2009) as it typically underscores individuals' contributions to research that benefit society as a whole. But it does more than that. In its in attempts to override the layman's association of *clinical* with *pain* and dispel the layperson's fears of clinical trials, such discourse focuses on the emotional as well as cognitive aspects of *clinical trials*. Only carefully-designed specialised corpora extending the range of contexts on which exemplification is based will guide dictionary writers and others to the typical patterns of use of basic medical terms in today's society and thus provide socially as well as medically relevant definitions. This means embracing less easily capturable connotations such as the affective values of fear-inducing words like *clinical*, *disease,* and *cancer* to name just a few.

Consumer is another word that merits special treatment as it helps define what constitutes medical discourse in today's society. While I have already mentioned the links between *patient* and *consumer,* their relationship requires further consideration as multiple intersections exist between the food system continuum and the healthcare-clinical research cline, one of which relates to the care required to ensure consumer protection in the form of food safety, which is determined through the analysis of specialised food system datasets, as underscored by the World Health Organisation (WHO):

> Information is required for food safety decision-making by all stakeholders in the food system continuum − from primary producers through to the consumer and all the actors in between, including risk assessors, policy-makers and communicators. Despite the increasing complexity of food systems, digital technologies are permitting the collection of an unprecedented amount of data from a virtually unlimited number of points along and around the food chain. The synthesis of these massive amounts of data requires considerable investment but can yield

> unparalleled insights and information applicable to food safety, public health and trade never before possible following the analysis of smaller isolated datasets. (WHO: Digitalization, Food Safety and Trade 2019, p. 1)

In the light of this statement with its focus on the significance of data management in links between trade, food safety and public health, it is hardly surprising that Sabrina Fusari's article carries out a thorough investigation of the key words used in one of the specific intersections between these two clines, namely the link between cancer and human consumption of meat. Nor is it surprising that her paper, and the specialised corpus she has created, both make multiple uses of the word *consumer*. Fusari's corpus is, in fact, mostly made up of responses by the scientific and medical community as well as international organisations to a document published by the *International Agency for Research on Cancer (IARC)*, the WHO's specialized cancer agency which promotes international cancer research collaboration. Acutely, Fusari's paper points out the need to understand that expressions in contemporary medical and scientific publications that look like everyday discourse often turn out to be specialised discourse with meanings quite different to the ones they might have been assumed to have. Thus, as Fusari points out, *strength of scientific evidence* is, as we have already seen from the discussion of EBM, a reference to the systems used to grade the quality of data in research, not a healthcare indication of the risks of eating too much meat. As Fusari puts it: "the intrinsic truthfulness of the IARC findings, or the extent to which they should revolutionize the public's eating habits to protect them against cancer risk, is beside the point: what matters is the rigour of the scientific analysis provided, as well as the soundness of its methodological approach".

As such, her paper adds to our understanding of the tense relationship between the specialised and the everyday in both healthcare and clinical research as the subject matter is inevitably a contentious Public Health issue. While terms such as *strength of scientific evidence* are well-known traps for the unwary, the process of defining terms technically and scientifically is far more deeply rooted than might at first be suspected. For example, the misalignment that Fusari quotes between what the *FAO/WHO Food Standards Programme* says counts as meat in its *Codex Alimentarius* and what the *U.S. Department of Health and Human Services* says in its *Dietary Guidelines for Americans* becomes more than comprehensible when we understand that they represent very different positions on the healthcare-clinical research cline. Thus, while the *Dietary Guidelines for Americans* documentation is part of a national consumer education programme urging individual consumers to adopt specific daily dietary habits, the *Codex Alimentarius* is instead a sixteen-volume compilation of general principles, general standards, definitions, codes, commodity standards, methods and recommendations published in English, French and Spanish addressing nations and their food safety policies.

Indeed Volume 10 deals with *Meat and meat products; soups and broths* and is the result of the work of the *Joint FAO/WHO Codex Alimentarius Commission* whose basic task is the preparation of food standards. No wonder then that the IARC's promotion of international cancer research collaboration is such an uphill struggle.

A first step towards implementing Fusari's own recommendation that the scientific community should accept the participation and intermediation of linguists in developing scientifically precise definitions and taxonomies for *meat* lies, in my view, precisely in the encouragement, manifested in all the papers in this volume, for discourse analysts to explore the process of consultation, negotiation and decision-making, whether carried out by governmental and inter-governmental institutions or by medical teams, and to determine the potential effects of their various positionings through specialised corpora. Just as we need to understand what *meat* means in different cultures, so we need to understand what *health* and *healthcare* means in contemporary Medicine in the different contexts and cultures in which these terms are used. In this respect, a more comprehensive study of how the food system continuum intersects with the healthcare-clinical research continuum would be a valuable starting point, as it would need to go beyond issues of data confidence and precision and deal instead with the need for food safety agencies to reflect on trust-building communication strategies for the poor as well as the rich:

> Food safety authorities should evaluate the best ways to harness new information and communication technologies to enhance consumer awareness and build trust, keeping in mind it is often difficult for consumers to differentiate between fact-based stories and unverified and false information. Additionally, it is important to recall that access to information via the internet is biased by wealth status, level of education, location (urban *vs.* rural) and gender. A focus on digital communication strategies could disadvantage segments of the society in need of particular attention with respect to food safety information. (WHO: Digitalization, Food Safety and Trade, 2019, p. 4)

In urging the construction of specialised corpora that explore the management of information in the field of food safety and its intersections with Public Health, it is, however, important to reflect once more on the role of technology, which, as we have already seen, is a likely source of change in the semantic profile of basic words associated with contemporary Medicine. Indeed, what is particularly interesting about the WHO's *Digitalization, Food Safety and Trade* publication is that its promotion of Big Data, IoT and artificial intelligence balances out their potential contributions to food safety for some communities with a need to understand their drawbacks for others, which means that key words such as *health hazard* and *risk assessment* need to be carefully tracked through specialised corpora vis-à-vis subtle changes in their meanings:

> Importantly, AI applications are being applied in the field of food safety risk assessment. Chemical risk assessments have traditionally relied on costly and time-consuming modelling based on animal testing, limiting throughput and raising animal welfare concerns and applicability to humans. With the current ability of computational and mathematical approaches using large quantities of data, predictive models are being generated that are based on high throughput cellular and *in vitro* assays, structural homology of chemicals and shared biochemical pathways, with the goal of facilitating a more inclusive risk assessment that ultimately is expected to aid in the faster and cheaper development of international food safety standards […] Machine learning is being employed to harness the wealth of foodborne pathogen genomic sequence data to predict health outcomes and improve hazard characterization of specific pathogens in risk assessment models. […] Use of such "black-box" techniques is problematic from both scientific and regulatory transparency perspectives; presents challenges for legal enforcement and communication and represents a potential barrier for adoption of the use of this technology. (WHO: Digitalization, Food Safety and Trade 2019, p. 2)

The paper by Stefania Maci and co-authors completes our survey of basic terms but also raises the issue of the interplay between general corpora and specialised corpora. I do not want to enter into the issue of what constitutes a specialised corpus or what constitutes a general corpus, at least as regards the issue of size since there is no theoretical reason why a specialised corpus could not be as big as, or even bigger than, the 100 million word *British National Corpus (BNC)* (source *www.natcorp.ox.ac.uk/corpus/*) that Maci and her co-authors use. Even if it was in some way pared down to include texts that prioritised food safety hazards, a food safety corpus would probably be regarded as a specialised corpus of immense proportions given the WHO's description of the domain in terms of zettabytes:

> Worldwide, over 25 billion devices are currently connected to the internet. Around the globe, the total number of sensors, monitors, computers, smartphones and other devices communicating with each other—through the Internet of Things (IoT)—is expected to exceed 75 billion by 2025. When applied to food safety, it is important to recognize that data may be collected from a very wide variety of sources and sectors (e.g. precision agriculture fertilization history, transport temperatures, geo-spatial, environmental and temporal metadata, hospitals records, ports of entry for imported foods, or sensors on individuals refrigerators or attached to personal smart phones). Such data complexity mirrors the increasing complexity of food supply chains and requires enormous (zettabytes) amounts of storage. Data mining tools such as web crawling, web scraping, data-mining and text extraction from scientific, industry and government databases can yield valuable information to better understand food safety hazards, and control measures and their implications for trade. (WHO: *Digitalization, Food Safety and Trade* 2019, p. 2)

The relationship between size and specialisation has, of course, been broached in the field of healthcare communication, for example, by Atkins and Harvey (2010) who refer to Sinclair (1991) and Flowerdew (2004), when describing the compilation of their corpus on young people's beliefs about health and illness:

> Although we ideally wanted to collect more data, taken by Sinclair's (1991: 18 pronouncement that a corpus should be as large as possible), we argue that, for the purpose of beginning to identify and describe patterns and commonalities in young people's beliefs about health and illness, one million words is a sufficient amount of data, or at the very least constitutes a

> substantial starting point. Given its size and focus on a particular communicative setting (the domain of adolescent health advice seeking), the corpus constitutes a specialised corpus. For a specialised corpus, one million words is by no means a small amount (according to Flowerdew 2004: 19, a corpus is generally considered small if it contains no more than 250000 words). […]. (Atkins, Harvey 2010, p. 608)

To my mind *specialised*, when applied to corpora, relates not to size but rather to the way in which a corpus is either constructed or used. Thus Atkins and Harvey have constructed a corpus that is specialised insofar as it relates to a specific *domain* (HIV/AIDS), a specific *age group* (adolescents), a specific *genre* (adolescents' health emails whose hallmark is the "frank and meticulous detail of their self-disclosures"), a specific *goal* (providing a socially and medically useful survey) and a specific *message* (the need for evidence-based sex education programmes). Moreover, what really counts is the result – the evidence-based nature of the corpus. The data on young people's (mis)understanding about health and HIV/AIDS was not elicited using traditional questionnaire-based methods and is thus all the more viable and valuable as a "unique vantage point from which to survey contemporary adolescent sexual health", one with "practical relevance for health practitioners and educators concerned with the health of young people" (Atkins, Harvey 2010, p. 616).

However, the article by Stefania Maci and her co-authors is also a demonstration that, within corpus studies, *specialised* may also refer to the uses to which much larger corpora such as the BNC can be put. Thanks to the *specialised* technique of semantic profiling which relates frequency counts of collocational patterns to semantic tags such as *BODY PART*, *TRANSMISSION* and *TREATMENT*, the *BNC* can be explored in terms of the typical contexts in which *illness*, *disease* and *sickness* outperform each other in terms of frequency. The result is that a much higher level of precision is reached in understanding the differences between these near synonyms when compared with definitions given by the various dictionaries that Maci and her co-authors consulted, whose shortcomings they describe very much along lines already illustrated above in relation to *clinical*. Significantly, what again counts in this approach is the *specific* application to which the emerging evidence can be put – in this case, its value in medical training in English for medical trainees whose L1 is not English and who have difficulty in grasping the typical contextual distributions of words like *illness*, *disease* and *sickness* as they are not likely to have precise matches in other languages. This is something that Maci and her co-authors successfully describe in relation to German and Italian but with the intention to "gain insights into potential translation problems of medical terms and phrases from English into other languages, for instance, Georgian, German, Italian, Hungarian and Polish, and vice versa."

The interplay between different corpora, as a confirmatory device for scholars in the pursuance of their research, has long been be pursued in corpus studies (Bianchi 2012, p. 52) and is, in part, facilitated by a third port of call: www.english-corpora.org/ which now hosts many corpora that were formerly located at https://corpus.byu.edu/, and which *inter alia* facilitates comparison of results obtained from specialised corpora with those of general corpora. However, such checking also needs to be seen in terms of applications in teaching and learning contexts in recognition of the fact that, as Maci and her co-authors state, a "clear understanding of the semantic profiles of the English terms will facilitate the selection of the most appropriate equivalent in any given context." Indeed, they rightly posit that the semantic profiling they have used could be applied to other corpora such as the *Corpus of Contemporary American English (COCA)*, for confirmatory evidence.

Potentially, semantic profiling is thus a tool which can indicate at what points on the healthcare-research cline these words occur relative to each other, something that a *domain-specific or genre-specific* corpus, by virtue of the restrictions that it has chosen to adopt, can normally only do with reference to one or two points on the cline. It is, of course true that:

> The methodological advantage of using a specialised corpus is that its smaller size lends itself to a more detailed, qualitative based examination than is possible with larger, more general corpora, such as their 100 million word British National Corpus or the Bank of English. The close examination of concordance lines with recourse to the linguistic co-text, for example, provides a rich source of data to complement more quantitative-based studies. (Atkins, Harvey 2010, p. 608)

However, medical trainees struggling to understand the difference between *sickness* and *illness* or indeed between *health*, *wellness* and *well-being* are unlikely to be happy with a qualitative-based examination that takes the form of a 'close examination of concordance lines' and will prefer far simpler visual takeaways such as a Venn diagram showing the degree to which these terms overlap in the different contexts along a cline, or other forms of simulation. Indeed, the fact that in the current volume there is a single surviving example of the once ubiquitous concordance and just a handful of references to type/token ratios – spot them if you can – might lead some to complain that this volume is not about *corpus linguistics* but also others to point to the usefulness of a volume of *corpus studies* that promotes cultural, social, technological and educational aspects when exploring medical texts and genres in English. Corpus interfaces need to provide syntheses of data from different sources such as dashboard combinations of various types of data from general and specialised corpora in keeping with the many types of displays and other innovative forms of data presentation now found in many clinical research sectors and many everyday healthcare contexts, all of which provide much-needed immediacy of interpretation. Even if this at the expense of

marginalising traditional concerns such as *POS tagging* and *mark-up* mentioned only briefly in this volume, such an approach ensures specialist knowledge really does meet up with everyday experience and needs.

3. Corpus as simulation

It follows from the previous section that a specialised corpus can be defined in relation to: (a) the texts it contains; (b) the units of analysis it presupposes; (c) the social uses and applications it permits. It also follows that specialised corpora can combine these characteristics in ways that, at first sight at least, may be viewed as surprising and which constitute a challenge to established conceptions of specialised corpora. This is the case with the two papers in this volume that refer to the *House Corpus* in which the primacy of words, if not questioned, is muted by the need for a more complete representation of medical events and interactions.

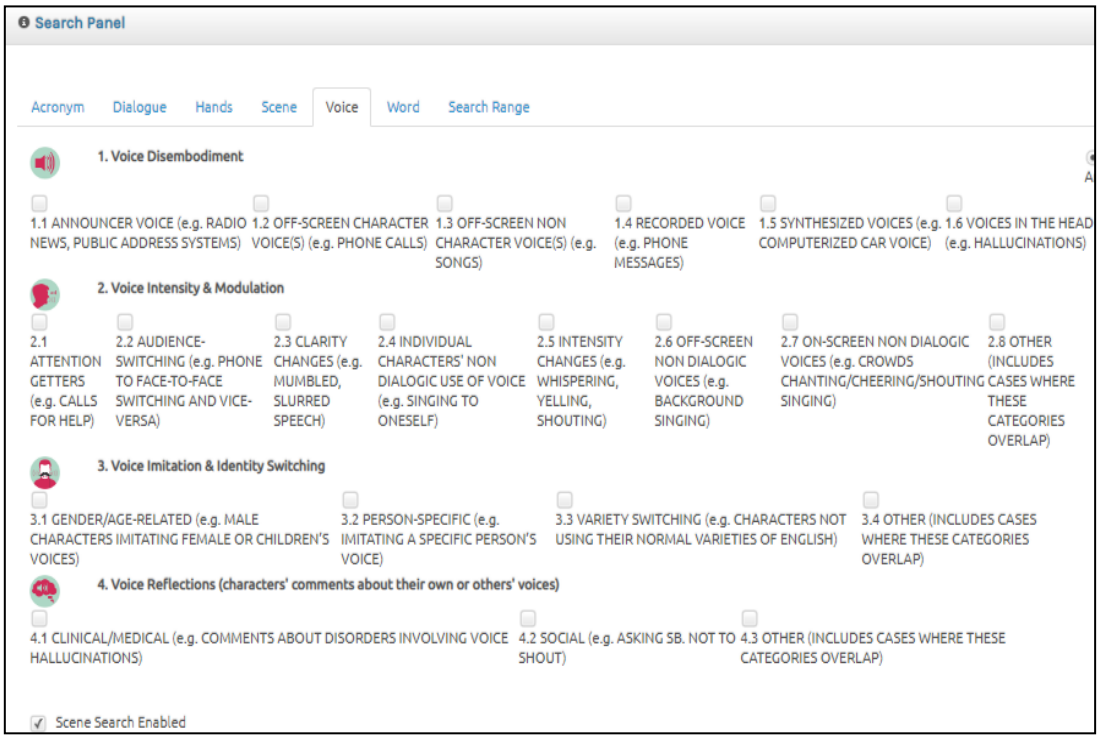


Figure 2
Screenshot of the *House Corpus* highlighting possible selections for Voice.

Figure 2 shows the search panel of this corpus which allows many different types of searches – linguistic, multimedia and multisemiotic – to be carried out, many of them as combinations of these search types. For example, a search might start out as purely linguistic in nature, looking for a specific lexical item's occurrences both in its (multi)word form and/or in its abbreviated

acronym form. However, this can become a multimedia search when associated to the *Scene Search Enabled* function, as this allows viewings of the scenes to be made in which the searched-for lexical items occur. However, yet a further step can be taken. As Figure 2 shows, a search in this specialised corpus can identify those scenes which include, for instance, expressions of voice intensity – shouting, whispering etc. – thus making the results more selective, both quantitively (fewer examples to discard) and qualitatively (greater specificity). The searched-for words thus come to be explictly characterised *multisemiotically*, in this case in terms of two interdependent but analytically and functionally different meaning-making resources: *voice quality* and *language*. This is in addition to being illustrated in the context of the scenes in which they occur, i.e. *multimedially*.

The potential and flexibility of this corpus is such that a search may omit lexical items altogether as searches can be implemented, for example, that identify those scenes where hallucinations occur (*SELECTION 1.6*) or those where voice disorders, such as 'voices in the head' are discussed. When, by means of the *SCENE* menu, selections are added that pinpoint activities carried out in specific hospital locations such as the ICU unit, the Maternity Ward or the Biopsy Room and then associate them with other selections referring to specialised hand movements using the *HANDS* menu, we can begin to see a new application for specialised medical corpora emerging, one that relates to the world of simulation in medical training, in this case a simulation of a hospital in terms of its activities and interactions. Thus, as the article by Davide Taibi, Ivana Marenzi and Qazi Asim Ijaz Ahma explains, the *House Corpus* has been constructed in such a way that it can easily be incorporated into advanced teaching and training activities, a matter further illustrated in the article by Anna Loiacono and Francesca Tursi in relation to medical trainees' learning about the abbreviatory strategies used in scientific discourse in English. Indeed, when using this corpus it becomes easier to show where the abbreviatory strategies used in English resemble those of other languages but, equally, how they also differ from them.

That instruction in medical discourse in English can be framed within simulations is an important step forward in terms of its integration into the training frameworks used vis-à-vis both undergraduate and postgraduate medical trainees. Investment in simulation in Medicine is growing by the day and takes many forms that run from mannequins to virtual reality:

> From the first "blue box" flight simulator to the military's impetus in the transfer of modeling and simulation technology to medicine, worldwide acceptance of simulation training is growing. Large collaborative simulation centers support the expectation of increases in multidisciplinary, interprofessional, and multimodal simulation training. Virtual worlds, both immersive and Web-based, are at the frontier of innovation in medical education. (Rosen 2008, p. 157)

Time and again, the issues of teamwork that I have described above are expressed in the medical literature in terms of failures in teamwork and communication with a solution being sought in simulation:

> Medical errors are one of the leading causes of death annually in the United States. Many of these errors are related to poor communication and/or lack of teamwork. Using simulation as a teaching modality provides a dual role in helping to reduce these errors. Thorough integration of clinical practice with teamwork and communication in a safe environment increases the likelihood of reducing the error rates in medicine. By allowing practitioners to make potential errors in a safe environment, such as simulation, these valuable lessons improve retention and will rarely be repeated. (Kuehster, Hall 2010, p. 123)

Only time will tell whether specialised corpora and medical simulation can meet up in ways that embrace some of the many activities and services that populate the healthcare-clinical research cline that I have sketched out above.

I wish to dedicate this Introduction, and indeed this volume, to the memory of Guy Aston, a pioneer in corpus linguistics. I had the fortune for brief periods in the 1970s to be his colleague both in Faculty of Letters, University of Bologna and in Pescara at the Libera Università Abruzzese, now Università degli Studi "G. d'Annunzio" Chieti-Pescara, and will never forget the courage he showed when facing up to the difficulties shared by all teachers of English linguistics in those demanding times; nor will I forget the great kindness he showed towards me personally on the few yet memorable occasions we met.

Bionote: Anthony Baldry's engagement with medical discourse began in December 1979 with a course he taught at the University of Pavia to postgraduate students which included reflection on the then recently published *Glasgow Coma Scale* and on the correspondence between Italian and English medical terminology vis-à-vis this and other key medical texts. Forty years later, after teaching hundreds of courses on medical discourse in English, in various Italian universities, he continues to engage with medical discourse with the same passion mostly within a sociolinguistic approach that explores the evolution of medical genres over time and which makes particular reference to multisemiotic corpora.

E-mail: anthony.baldry@gmail.com

References

Aathira R. and Jain V. 2014, *Advances in management of type 1 diabetes mellitus*, in "World journal of diabetes" 5 [5], pp. 689-696.

Agarwal S. and Lau C.T. 2010, *Remote health monitoring using mobile phones and Web services*, in "Telemedicine and e-Health" 16 [5], pp. 603-607.

Ahmed Z., Garfield S., Jani Y., Jheeta S., Franklin B. and Barber N. 2016, *Impact of electronic prescribing on patient safety in hospitals: Implications for the UK*, in "Clinical Pharmacist" 8, pp. 1758-9061.

Alberich-Bayarri A. 2017, *Image interpretation*, in Donoso-Bach L. and Boland G.W.L. (eds.), *Quality and Safety in Imaging*, Springer, Cham, Switzerland, pp. 135-143.

Atkins S. and Harvey K. 2010, *How to use corpus linguistics in the study of health communication*, in O'Keeffe A. and McCarthy M. (eds), *The Routledge handbook of corpus linguistics*, Routledge, London/New York, pp. 633-647.

Baldry A.P. 2005, *A multimodal approach to text studies in English*. Palladino, Campobasso.

Baldry A.P. 2011, *Multimodal web genres: Exploring scientific English*, Ibis, Como/Pavia.

Baldry A. and Kantz D. 2009, *New dawns and new identities for multimodality. Public information films in the national archives*, in Vasta N. and Rosa Caldas-Coulthard C. (eds), *Identity Construction and Positioning in Discourse and Society*, "Textus" 22 [1], pp. 225-255.

Beard R.J. 2017, *Electronic Prescribing and Robotic Dispensing: The Impact of Integrating Together on Practice and Professionalism*, in Dekoulis G. (ed.), *Robotics: Legal, Ethical and Socioeconomic Impacts*, InTech open, Rijeka, Croatia, pp. 133-152.

Bewley W.L. and O'Neil H.F. 2013, *Evaluation of medical simulations*, in "Military medicine" 178 [suppl_10], pp. 64-75.

Bianchi F. 2012, *Chapter 3 - Corpora and corpus linguistics*, in Bianchi F., *Culture, corpora and semantics. Methodological issues in using elicited and corpus data for cultural comparison*, ESE – Salento University Publishing, Lecce, pp. 31-54. http://siba-ese.unisalento.it/index.php/culturecorpora/article/view/12434/11073 (10.07.2019).

Breton M., Farret A., Bruttomesso D., Anderson S., Magni L., Patek S., Dalla Man C., Place J., Demartini S., Del Favero S. and Toffanin C. 2012, *Fully integrated artificial pancreas in type 1 diabetes: modular closed-loop glucose control maintains near normoglycemia*, in "Diabetes" 61 [9], pp. 2230-2237.

Chalmers I., Milne I., Tröhler U., Vandenbroucke J., Morabia A., Tait G. and Dukan E. 2008, *The James Lind Library: explaining and illustrating the evolution of fair tests of medical treatments*, in "The journal of the Royal College of Physicians of Edinburgh" 38 [3], pp. 259-264.

Chiauzzi E., Rodarte C. and DasMahapatra P. 2015, *Patient-centered activity monitoring in the self-management of chronic health conditions*, in "BMC medicine" 13. https://bmcmedicine.biomedcentral.com/track/pdf/10.1186/s12916-015-0319-2 (24.10.2019).

Churchouse C. and McCafferty C. 2012, *Standardized patients versus simulated patients: is there a difference?*, in "Clinical Simulation in Nursing" 8 [8], pp. e363-e365.

Clarke W.L., Anderson S., Breton M., Patek S., Kashmer L. and Kovatchev B. 2009, *Closed-loop artificial pancreas using subcutaneous glucose sensing and insulin delivery and a model predictive control algorithm: the Virginia experience*, in

"Journal of Diabetes Science and Technology" 3 [5], pp. 1031-1038. https://journals.sagepub.com/doi/pdf/10.1177/193229680900300506 (23.10.2019).

Coleman E.A. 2003, *Falling through the cracks: challenges and opportunities for improving transitional care for persons with continuous complex care needs*, in Journal of the American Geriatrics Society" 51 [4], pp. 549-555.

Conti A.A. and Gensini G.F. 2008, *Doctor-patient communication: a historical overview*, in "Minerva medica" 99 [4], pp. 411-415.

Coons M.J., Greiver M., Aliarzadeh B. *et al*. 2017, *Is glycemia control in Canadians with diabetes individualized? A cross-sectional observational study*, in "BMJ Open Diabetes Research & Care" 5 [1]. https://drc.bmj.com/content/bmjdrc/5/1/e000316.full.pdf (24.10.2019).

Crawford P., Brown B. and Harvey K. 2014, *Corpus linguistics and evidence-based health communication*, in Hamilton H. and Chou W-y.S. (eds.), *The Routledge Handbook of Language and Health Communication*, Routledge, London/New York pp. 75-90.

Crawford S.Y., Grussing P.G., Clark T.G. and Rice J.A. 1998, *Staff attitudes about the use of robots in pharmacy before implementation of a robotic dispensing system*, in "American journal of health-system pharmacy" 55 [18], pp. 1907-1914.

Doyle F.J., Huyett L.M., Lee J.B., Zisser H.C. and Dassau E. 2014, *Closed-loop artificial pancreas systems: engineering the algorithms*, in "Diabetes care" 37 [5], pp. 1191-1197.

Evans W.E. and Relling M.V. 2004, *Moving towards individualized medicine with pharmacogenomics*, in "Nature" 429, pp. 464-468.

Ferguson G. 2001, *If you pop over there: a corpus-based study of conditionals in medical discourse*, in "English for Specific Purposes" 20 [1], pp. 61-82.

Flowerdew L. 2004, *The argument for using English specialized corpora to understand academic and professional language*, in Connor U. and Upton T.A. (eds.), *Discourse in the professions: Perspectives from corpus linguistics*, John Benjamins, Amsterdam/Philadelphia, pp. 11-33.

Goebel A. 2011, *Complex regional pain syndrome in adults*, in "Rheumatology" 50 [10], pp. 1739-1750.

Gonçalves-Bradley D.C., Lannin N.A., Clemson L.M., Cameron I.D., Shepperd S. 2016, *Discharge planning from hospital*, in "Cochrane Database of Systematic Reviews" https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD000313.pub5/epdf/full (24.10.2019).

Gorini A., Gaggioli A. and Riva G. 2008, *A second life for eHealth: prospects for the use of 3-D virtual worlds in clinical psychology*, in "Journal of medical Internet research" 10 [3]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2553247/ (24.10.2019).

Goundrey-Smith S.J. 2019, *Technologies that transform: digital solutions for optimising medicines use in the NHS*, in "BMJ Health & Care Informatics" 26 [1]. https://informatics.bmj.com/content/26/1/e100016 (24.10.2019).

Guyatt G.H., Oxman A.D., Vist G.E., Kunz R., Falck-Ytter Y., Alonso-Coello P. and Schünemann H.J. 2008, *GRADE: an emerging consensus on rating quality of evidence and strength of recommendations*, in "BMJ" 336 [7650], pp. 924-926.

Harrison N. 2018, *Regressing or progressing: what next for the doctor-patient relationship?*, in "The Lancet Respiratory Medicine" 6 [3], pp. 178-180.

Heinrichs W.L., Youngblood P., Harter P.M. and Dev P. 2008, *Simulation for team training and assessment: case studies of online training with virtual worlds*, in "World journal of surgery" 32 [2], pp. 161-170.

Heritage J. and Maynard D.W. (eds.) 2006, *Communication in medical care: Interaction between primary care physicians and patients*, Vol. 20, Cambridge University Press, Cambridge.

Hopkins A. and Dudley-Evans T. 1988, *A genre-based investigation of the discussion sections in articles and dissertations*, in "English for specific purposes" 7 [2], pp. 113-121.

Hunt D. and Harvey K. 2015, *Health communication and corpus linguistics: using corpus tools to analyse eating disorder discourse online*, in Baker P. and McEnery T (eds.), *Corpora and Discourse Studies. Integrating Discourse and Corpora*, Palgrave, London, pp. 134-154.

Hyland K. 1998, *Hedging in scientific research articles*, John Benjamins, Amsterdam/Philadelphia.

Ienca M., Fabrice J., Elger B., Caon M., Pappagallo A.S., Kressig R.W. and Wangmo T. 2017, *Intelligent assistive technology for Alzheimer's disease and other dementias: a systematic review*, in "Journal of Alzheimer's Disease" 56 [4], pp. 1301-1340.

Istepanian R.S., Al-Anzi T. 2018, *m-Health 2.0: new perspectives on mobile health, machine learning and big data analytics*, in "Methods" 151, pp. 34-40.

Jameson J.L. and Longo D.L. 2015, *Precision medicine – personalized, problematic, and promising*, in "Obstetrical & gynecological survey" 70 [10], pp. 612-614.

Johansson D., Malmgren K. and Murphy M.A. 2018, *Wearable sensors for clinical applications in epilepsy, Parkinson's disease, and stroke: a mixed-methods systematic review*, in "Journal of neurology" 265 [8], pp. 1740-1752.

Kaba R. and Sooriakumaran P. 2007, *The evolution of the doctor-patient relationship*, in "International Journal of Surgery" 5 [1], pp. 57-65.

Keating E. and Mirus G. 2003, *American Sign Language in virtual space: Interactions between deaf users of computer-mediated video communication and the impact of technology on language practices*, in "Language in Society" 32 [5], pp. 693-714.

Kononowicz A.A., Zary N., Edelbring S., Corral J. and Hege I. 2015, *Virtual patients-what are we talking about? A framework to classify the meanings of the term in healthcare education*, in "BMC medical education" 15 [1], pp. 11-18.

Kontos N., Querques J. and Freudenreich O. 2011, *Two's company, three hundred million's a crowd: balancing clinical integrity and population consciousness in medical education*, in "Academic Medicine" 86 [11], p. 1341.

Kourtis L.C., Regele O.B., Wright J.M. and Jones G.B. 2019, *Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity*, in "NPJ digital medicine" 2 [9]. https://www.nature.com/articles/s41746-019-0084-2.pdf (24.10.2019).

Kress G. 1998, *Visual and verbal modes of representation in electronically mediated communication: The potentials of new forms of text*, in Snyder I. (ed.), *Page to screen: Taking literacy into the electronic era*, Routledge, London/New York, pp. 53-79.

Kuehster C.R. and Hall C.D. 2010, *Simulation: Learning from mistakes while building communication and teamwork*, in "Journal for Nurses in Professional Development" 26 [3], pp. 123-127.

Lanzola G., Losiouk E., Del Favero S., Facchinetti A., Galderisi A., Quaglini S. and Cobelli C. 2016, *Remote blood glucose monitoring in mHealth scenarios: A review*, in "Sensors" 16 [12], pp. 1983- 2000.

Laplante P.A. and Laplante N. 2016, *The internet of things in healthcare: Potential applications and challenges*, in "IT Professional" 18 [3], pp. 2-4.

Levine A.I., DeMaria Jr S., Schwartz A.D. and Sim A.J. (eds.) 2013, *The comprehensive textbook of healthcare simulation*, Springer Science & Business Media, New York.

MacDonald M.N. 2002, *Pedagogy, pathology and ideology: The production, transmission and reproduction of medical discourse*, in "Discourse & Society" 13 [4], pp. 447-467.

Mamatas N.I.C.K. 2007, *Why we love Holmes and love to hate House*, in Wilson L. (ed.), *HOUSE unauthorized. Vasculitis, Clinic Duty, and Bad Bedside Manner*, Benbella Books Inc., Dallas, pp. 87-96.

Mason J.B. 2008, *The new demands by patients in the modern era of total joint arthroplasty*, in "Clinical orthopaedics and related research" 466 [1], pp. 146-152.

Maynard D.W. and Heritage J. 2005, *Conversation analysis, doctor-patient interaction and medical communication*, in "Medical education" 39 [4], pp. 428-435.

McGrath J.L., Taekman J.M., Dev P., Danforth D.R., Mohan D., Kman N., Talbot T.B. 2018, *Using virtual reality simulation environments to assess competence for emergency medicine learners*, in "Academic Emergency Medicine" 25 [2], pp. 186-195.

Menz F. and Al-Roubaie A. 2008, *Interruptions, status and gender in medical interviews: The harder you brake, the longer it takes*, in "Discourse & Society" 19 [5], pp. 645-666.

Metcalf D., Milliard S.T., Gomez M. and Schwartz M. 2016, *Wearables and the internet of things for health: Wearable, interconnected devices promise more efficient and comprehensive health care*, in "IEEE pulse" 7 [5], pp. 35-39.

Milani R.V. and Lavie C.J. 2015, *Health care 2020: reengineering health care delivery to combat chronic disease*, in "The American journal of medicine" 128 [4], pp. 337-343.

Morris G.H. and Chenail R.J. 2013, *The talk of the clinic: Explorations in the analysis of medical and therapeutic discourse*, Routledge, London/New York.

Murad M.H., Asi N., Alsawas M. and Alahdab F. 2016, *New evidence pyramid*, in "BMJ Evidence-Based Medicine" 21 [4], pp. 125-127.

Naish J. 2018, *What is the truth on hormone replacement therapy*, in "Daily Mail". www.dailymail.co.uk/health/article-6128001/What-truth-hormone-replacement-therapy.html (24.10.2019).

Naylor M.D., Brooten D., Campbell R., Jacobsen B.S., Mezey M.D., Pauly M.V. and Schwartz J.S. 1999, *Comprehensive discharge planning and home follow-up of hospitalized elders: a randomized clinical trial*, in "Jama" 281 [7], pp. 613-620.

O'Keeffe A. and McCarthy M. (eds.) 2010, *The Routledge handbook of corpus linguistics*, Routledge, London/New York.

Ong L.M., De Haes J.C., Hoos A.M. and Lammes F.B. 1995, *Doctor-patient communication: a review of the literature*, in "Social science & medicine" 40 [7], pp. 903-918.

Ozanne A., Johansson D., Hällgren Graneheim U., Malmgren K., Bergquist F. and Alt Murphy M. 2018, *Wearables in epilepsy and Parkinson's disease – A focus group study*, in "Acta Neurologica Scandinavica" 137 [2], pp. 188-194.

Parenti N., Reggiani M.L.B., Iannone P., Percudani D. and Dowding D. 2014, *A systematic review on the validity and reliability of an emergency department triage scale, the Manchester Triage System*", in "International journal of nursing studies" 51 [7], pp. 1062-1069.

Pearce W. and Raman S. 2014, *The new randomised controlled trials (RCT) movement in public policy: challenges of epistemic governance*, in "Policy sciences" 47 [4], pp. 387-402.

Picard R. 2014, *November. Affective media and wearables: surprising findings,* in *Proceedings of the 22nd ACM international conference on Multimedia,* ACM, pp. 3-4. https://dl.acm.org/citation.cfm?id=2647959 (14.10.2019).

Pizzini F. (ed.) 1990, *Asimmetrie comunicative: differenze di genere nell'interazione medico-paziente*, Vol. 14, Franco Angeli, Milano.

Pronovost P. and Vohr E. 2010, *Safe patients, smart hospitals. How one doctor's checklist can help us change health care from the inside out*, Penguin, London.

Reese C.E., Jeffries P.R. and Engum S.A. 2010, *Learning together: Using simulations to develop nursing and medical student collaboration*, in "Nursing education perspectives" 31 [1], pp. 33-37.

Rizzo A.A. and Talbot T. 2016, *Virtual reality standardized patients for clinical training*, in Combs C.D., Sokolowski J.A. and Banks C.M. (eds.), *The digital patient: Advancing Healthcare, Research, and Education*, Wiley, Hoboken, NJ, pp. 257-272.

Rodbard D. 2016, *Continuous glucose monitoring: a review of successes, challenges, and opportunities*, "Diabetes technology & therapeutics" 18 [S2], pp. 3-13.

Rojahn K., LaplanteS., Sloand J., Main C., Ibrahim A., Wild J., Sturt N., Areteou T. and Johnson K.I. 2016, *Remote monitoring of chronic diseases: a landscape assessment of policies in four European countries*, in "PloS one" 11 [5], pp. 1-15. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0155738 (24.10.2019).

Rosen K.R. 2008, *The history of medical simulation*, in "Journal of critical care" 23 [2], pp. 157-166.

Rosner A.L. 2012, *Evidence-based medicine: revisiting the pyramid of priorities*, in "Journal of Bodywork and Movement Therapies" 16 [1], pp. 42-49.

Salager-Meyer F. 1991, *Medical English abstracts: How well are they structured?*, in "Journal of the American Society for Information Science" 42 [7], pp. 528-531.

Saucier A.N., Ansa B., Coffin J., Akhtar M., Miller A. *et al*. 2017, *Patient perspectives of an individualized diabetes care management plan*, in "European journal for person centered healthcare" 5 [2], pp. 213-219.

Schegloff E.A. 1999, *Discourse, pragmatics, conversation, analysis*, in "Discourse studies" 1 [4], pp. 405-435.

Schwartz L. 2002, *Is there an advocate in the house? The role of health care professionals in patient advocacy*, in "Journal of medical ethics" 28 [1], pp. 37-40.

Shah H., Rossen B., Lok B., Londino D., Lind S.D. and Foster A. 2012, *Interactive virtual-patient scenarios: an evolving tool in psychiatric education*, in "Academic Psychiatry" 36 [2], pp. 146-150.

Shaneyfelt T. 2016, *Pyramids are guides not rules: the evolution of the evidence pyramid*, in "BMJ Evidence-Base Medicine" 21 [4], pp. 121-122.

Shoeb M., Merel S.E., Jackson M.B. and Anawalt B.D. 2012, *"Can we just stop and talk?" patients value verbal communication about discharge care plans*, in "Journal of hospital medicine" 7 [6], pp. 504-507.

Silbert S. 2019, *All the Things You Can Track With Wearables: Steps and calories are just the tip of the iceberg.* https://www.lifewire.com/what-wearables-can-track-4121040 (24.10.2019).

Silverman D. 1987, *Communication and medical practice: Social relations in the clinic*, Sage, London.

Sinclair J. 1991, *Corpus, concordance, collocation*, Oxford University Press, Oxford.

Steele D.J., Jackson T.C. and Gutmann M.C. 1990, *Have you been taking your pills?: the adherence-monitoring sequence in the medical interview*, in "Journal of Family Practice" 30 [3], pp. 294-300.

Sur R.L. and Dahm P. 2011, *History of evidence-based medicine*, in "Indian journal of urology" 27 [4], pp. 487-489.

Surendran D., Janet J., Prabha D. and Anisha E. 2018, *A Study on devices for assisting Alzheimer patients*, in *2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, IEEE, pp. 620-625. https://ieeexplore.ieee.org/document/8653658 (24.10.2019).

Swales J. 1990, *Genre analysis: English in academic and research settings*, Cambridge University Press, Cambridge.

Swan M. 2009, *Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking*, in "International journal of environmental research and public health" 6 [2], pp. 492-525.

Trickett S.B. and Trafton J.G. 2007, *"What if…": The use of conceptual simulations in scientific reasoning*, in "Cognitive Science" 31 [5], pp. 843-875.

Ventola C.L. 2014, *Mobile devices and apps for health care professionals: uses and benefits*, in "Pharmacy and Therapeutics" 39 [5], 356-364.

Walker H.K. 1990, *The origins of the history and physical examination*, in Walker H.K., Hall W.D., Hurst J.W. (eds.), *Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition*, Butterworths, Boston.

Wieringa S., Engebretsen E., Heggen K. and Greenhalgh T. 2018, *Rethinking bias and truth in evidence-based health care*, in "Journal of evaluation in clinical practice" 24 [5], pp. 930-938.

White Jr K.P. 2005, *December. A survey of data resources for simulating patient flows in healthcare delivery systems*, in Kuhl M.E., Steiger N.M., Armstrong F.B. and Joines J.A. (eds.), *Proceedings of the 37th conference on Winter simulation*, pp. 926-935. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.2965&rep=rep1&type=pdf (24.10.2019).

World Health Organisation 2019, *Digitalization, food safety and trade*. https://www.who.int/docs/default-source/resources/digitalization-food-safety-and-trade-en.pdf?sfvrsn=a11a03b8_2 (14.10.2019).

Yuehong Y.I.N., Zeng Y., Chen X. and Fan Y. 2016, *The internet of things in healthcare: An overview*, in "Journal of Industrial Information Integration" 1, pp. 3-13.

Zabielska M. 2015, *Patient-centred case reporting: state of the art*, in Zabielska M., Wąsikiewicz-Firlej E. and Szczepaniak-Kozak A. (eds.), *Discourses in co (n) text: the many faces of specialised discourse*, Cambridge Scholars Publishing, Newcastle upon Tyne, pp. 2-32.

Zimerman A.L., 2013. *Evidence-based medicine: a short history of a modern medical movement*, in "AMA Journal of Ethics" 15 [1], pp.71-76.

HIV DISCOURSE IN THE *BRITISH MEDICAL JOURNAL*, 1985-2005 The Impact of digital literacy and Evidence-Based Medicine on syntactic patterns and variations in RA titles[1]

STEFANIA CONSONNI

Abstract – Although titling is traditionally a lexically and textually prominent operation, performing key informative/persuasive/promotional functions in discourse domains such as advertising and entertainment, the spread of Web-based communication has increased its importance with respect to practices farther away on a discoursal spectrum from such functions as medical communication. The inception of the Internet as the main channel for knowledge dissemination has brought about significant changes in the titling of highly specialized discourse. Medical RA titles (RATs) seem, as a genre, to provide insights into the impact of digital literacy on scientific knowledge. In order to explore such changes, a total of 1250 RATs from the *British Medical Journal* – the world's first online medical journal – were collected from a 20-year period, and analysed with AntConc and Wordsmith Tools. The RATs in the corpus trace the history of the Human Immunodeficiency Virus from 1985, when the first WHO conference on AIDS was held in the USA, until 2005. The paper analyses and contrasts print *vs*. digital RATs, identifying and quantifying the key syntactical/textual patterns and variations in a genre whose main function is to package/textualize scientific contents (including competing clinical methodologies), as well as to disseminate them across specialized and/or lay audiences. Research questions concern the extent to which the language of RATs has been changing with respect to the dissemination triggered by digital literacy, from crystallised and gate-keeping formulations to more articulated ones, placing distinctive emphasis on argumentative/persuasive/metadiscoursive functions, as well as the impact of Evidence-Based Medicine – today's leading paradigm for scientific knowledge, first presented in *BMJ* in 1995 – on contemporary HIV discourse.

Keywords: Medical titles; digital literacy; discourse analysis; HIV; Evidence-Based Medicine.

[1] This study is part of a national research project on "Knowledge Dissemination across media in English: Continuity and change in discourse strategies, ideologies, and epistemologies", financed by the Italian Ministry of University and Research (PRIN 2015TJ8ZAS).

Doi Code: 10.1285/i9788883051531p47

1. Introduction

This chapter analyses HIV discourse in the *British Medical Journal* in a time span of twenty years, from 1985 – when the first world conference on AIDS was held in the USA – to 1995, the year *BMJ* started to implement Evidence-Based Medicine (EBM) and to be published online, and from 1995 to 2005. RA titles (henceforth RATs) will be investigated as a key strategy for knowledge dissemination, by comparing their functions and impact before and after the inception of EBM and of digital literacy practices. The purpose of the study is to identify and quantify the key patterns and variations in a genre whose main function is to package/textualize scientific contents and to contribute to their widest possible dissemination, and thereby to explore the impact of new research procedures and new communication paradigms on the traditionally codified discourse of clinical knowledge. Insights will also be provided as to the linguistic history, in terms of both clinical representation and discursive dissemination, of a life-threatening and socially sensitive pathology.

The epistemological framework for this paper is provided in two classics on scientific expository practices:

1. In *Naissance de la clinique*, Michel Foucault (1963) argues that clinical knowledge was born at the end of the Eighteenth century as the truth effect of discourse practices producing a system of beliefs around the physiology and pathologies of the human body. Bodies, tissues and diseases entered the field of scientific truth, which is always framed within a specific discursive period: clinical authority relies on its relationship to the current organisation of knowledge, not so much to a non-discursive state of affairs (i.e., clinical reality as it is). Scientific truth is the result of ongoing negotiation between knowledge production and popularization, which explains why medical discourse has recently been evidenced as a contingent construction, varying among different periods and epistemologies, as well as across different pragmatic contexts.
2. As Shinn and Whitley (1985) argue, scientific discourse practices are ideologically non-neutral. Far from being "polished, objectified, linear and persuasive" (Bucchi 1998), scientific research depends on dissemination, a transactional phenomenon impacting on research in ways which cannot be detached from research itself, and involving a variety of actors and audiences. Clinical legitimization comes from audiences including not only fellow physicians and training experts, but also non-scientific audience segments (i.e., a number of professions drawing credibility from the use of scientific knowledge), as well as the growing business/corporate public (which may in turn seek legitimization from scientific discourse, while exerting influence on the purposes and directions of research), and the lay

public of popularization. Feedback from all the strata involved in this process produces and validates knowledge, and contributes to fixing research agendas throughout disciplines, especially in the case of socially impactful pathologies such as HIV.

As a matter of fact, the dissemination process inherent to medical expository practices has been immensely amplified over the last two decades by the Internet, that is, by the digital environment and Web-based communication strategies. In this respect, medical RATs have proved to be a crucial genre. Although titling has been – since classical rhetoric – a *per se* lexically, syntactically and textually prominent operation, one that typically performs key informative/persuasive/promotional functions in discourse domains such as the media, advertising and entertainment (Hartley 2005a, 2005b; Martin 1998; Straumann 1935), the spread of Web-based communication has increased its importance with respect to practices traditionally farther away on a discoursal spectrum from such functions, such as medical communication (Calsamiglia 2003; Calsamiglia, Van Dijk 2004; Jaime Sisó 2009; Giannoni 2014; Gotti *et al.* 2015; Myers 2003; Smith 2000; Soler 2007; Swales 2003).

By "medical RAs" this paper refers to specialized texts, generally aimed at a specialized audience of fellow researchers/clinicians, displaying the IMRD format (i.e., Introduction, Method-Materials, Results, Discussion, which all "evidence a good deal of experimental work"), and forming the genre which serves as a "generator of new knowledge about a specific subject" (Soler 2007, p. 92), and whose main expected pragmatic function is referential/informative. By "RATs" this paper refers to typically concise structures, preceding and associated to a longer text, which they both synthesize (in terms of informative content) and present in an efficient/appealing way (that is, providing accurate directions as regards the RA's text type and pragmatics). In medical communication, RATs can be said to perform a number of pragmatic functions:

1. Informativity: in its conciseness, transparency and completeness, the science title is "an up-front, straightforward presentation of information, whether the information is that of what the paper has established or what the paper is about" (Haggan 2004, p. 313). In terms of cognitive psychology, titles are advanced textual organizers, revealing preview information from a later, more extended text (Kozminsky 1977).
2. Retrievability of RAs in terms of online search engine optimization: "titles in publications are key elements in the organization and retrieval of scholarly data" (Soler 2007, p. 91), surrogating the document "in bibliographies, databases, indexes and reference lists" and the Web in general (Yitzhaki 1997, p. 220).

3. Attractiveness: the title attracts a reader's attention to a paper and presents its content from a short glimpse, "thus contributing to its initial selection or rejection" by other researchers (Hjørland, Nielsen 2001, p. 264).[2]

The present analysis considers RATs on HIV published in the *British Medical Journal* from 1985 to 2005. The choice of journal, as well as of time span and clinical specialty, is not unfounded, for two reasons:

1. In 1995, *BMJ* was the world's first general medical journal to go fully online.[3] The first research question of this paper therefore concerns the extent to which the language of RATs has been changing with respect to the global knowledge dissemination process brought about by electronic literacy, and more specifically, the extent to which such process has also been influencing the titling of highly specialized, expert-to-expert discourse, from markedly standardised, crystallised and gate-keeping formulations to more articulated textual, metadiscursive and pragmatic functions (Garzone 2006; Gotti 2003, 2013; Hyland 2005).
2. In 1995, *BMJ* started to systematically implement Evidence-Based Medicine, the most influential definition of which is provided by Sackett *et al.* (1996) in *BMJ* itself. EBM is today's leading paradigm for medical knowledge, first introduced in 1992 to set out completely new methodological procedures and protocols in the life sciences. EBM is "the use of mathematical estimates of the risk of benefit and harm, derived from high-quality research on population samples, to inform clinical decision-making in the diagnosis, investigation or management of individual patients" (Greenhalgh 2010, p. 1). It stands in opposition to traditional practice, which revolved around individual clinical expertise (the commonest approach until the early 1990s), in that it stems from "the best available external clinical evidence from systematic research" (Sackett *et al.* 1996, p. 71), i.e., from the systematic statistical analysis of data, which leads to the formulation of questions and testing of hypotheses.

EBM is based upon what is commonly referred to as the "pyramid of evidence", where several levels of evidence provided by clinical research are ranked according to their reliability. The levels are arranged in a system

[2] A summary of these three functions (Genette 1988, pp. 178-179) is provided in Zeiger (1991, cited in Wang, Bai 2005, p. 390): "the hallmarks of a good title are that it accurately, completely and specifically identifies the main topic or the main point of the paper, is unambiguous, is concise, and [provides] important term[s]" with reference to the clinical topic and/or the methodology/research protocols employed.

[3] Founded in 1840 as the *Provincial Medical and Surgical Journal*, the journal launched several medical discoveries of the Twentieth century, including the use of chloroform during Queen Victoria's eighth childbirth (1847), Joseph Lister's observations on antisepsis in surgery (1867-79), the link between Anopheles mosquito and malaria (1898), the first streptomycin trial (1948), and the first report on smoking and lung cancer (1950).

accounting for the strength of their results on the basis of the study design, i.e., the methodological description – involving participants, implements and procedures, as shown in Figure 1 – to be found in the Method section of RAs.

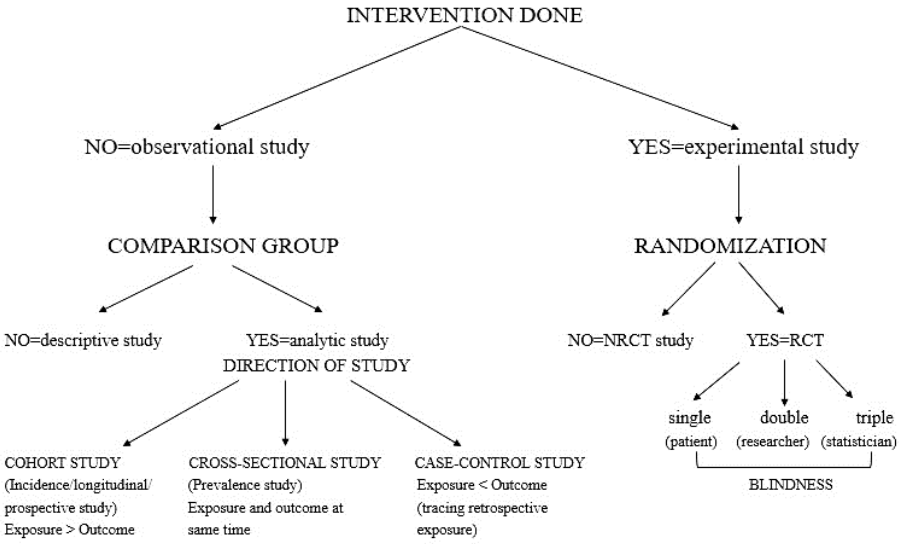


Figure 1
EBM study designs.

The Pyramid reveals how to weigh different levels of evidence in order to make health-related decisions (Greenhalgh 2010, pp. 18-45), putting the results of each study design into a hierarchy based on the relative strengths and weaknesses of each piece of research, as can be seen in Figure 2:

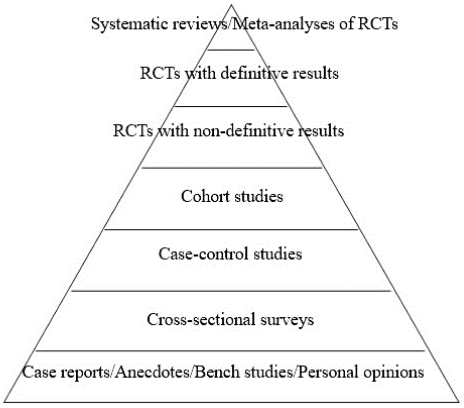


Figure 2
The pyramid of evidence.

Each level represents a different study design and corresponds to increasing quality/reliability of evidence and expected result accuracy, as well as decreasing chance of statistical error, and to minimized bias from confounding variables potentially influencing clinical results:

i. Systematic reviews of RCTs are gold-standard sources; started in the 1980s under the inspiration of Archibald Cochrane, they search broadly for clinical trials on a topic and pool the results statistically; they confront different findings among different studies on the same topic, which makes them likely to be robust and generalizable.
ii. Randomized Controlled Trials (RCTs) randomly allocate participants to either one intervention (e.g. drug treatment) or another (e.g. placebo treatment). Both groups are followed up for a specific period of time, and analysed in terms of specific outcomes defined at the outset of the study (e.g. death, heart attack, etc.). There can be several levels of blindness in an RCT, when patients, researchers and statisticians themselves are not informed as to how patients are allocated to interventions.
iii. In a cohort (longitudinal/incidence) study, a fixed sample of population is measured repeatedly on the same variables, providing a series of pictures illustrating change over time.
iv. In case control studies, patients undergo controls on past exposure to a possible causal agent for a particular condition (frequently used to determine the aetiology of disease, not treatment, e.g. rare conditions).
v. In cross-sectional (prevalence) surveys, a collection of information is taken only once from a given sample of population.
vi. Case reports are descriptions of a patient's medical history in the form of a story, and lie at the bottom of the pyramid with traditional forms of knowledge such as anecdotes, bench studies and personal opinions.

In the light of the above, the second research question in this paper takes into account the impact of EBM – as the gold-standard paradigm in scientific production and dissemination – on the language of medical RATs, and the changes in pragmatic scope and methodological positioning it brings about in contemporary medical literature on HIV.

It is also worth mentioning that the clinical specialty investigated in this paper is HIV, whose literary history in the international scientific community started exactly in 1985.[4] In March 1985 the FDA licensed the first ELISA commercial test to detect antibodies to the virus. In April the same year, the first WHO conference on AIDS was held in Atlanta, Georgia. In May 1985,

[4] The earliest case of infection with HIV-1 in a human was detected in 1959 in Congo. HIV-1 may apparently have originated in the 1940s or early 1950s. In the mid-1970s, the virus spread in the USA, where a number of cases of pneumonia, cancer and other pathologies were reported by doctors in LA and NY to be related to male homosexuality. In 1982 the term AIDS was first used to describe opportunistic infections and other pathologies linked to the virus. In 1983 the virus triggering AIDS was discovered; it was first named HTLV.III/LAV. The name was changed to HIV in 1986. In 1999 the origin of the HIV-1 virus in a subspecies of chimpanzees in west Africa was discovered; the first humans might have been infected by the animals' blood while hunting.

the International Committee on Taxonomy of Viruses ruled that the pathogen responsible for AIDS – first discovered in May 1983 by a French research team as a retrovirus called LAV – should be named the Human Immunodeficiency Virus.[5]

2. Materials and Method

For the purpose of this analysis,[6] a corpus of RATs has been assembled, covering the totality of RAs published in *BMJ* between 1985 and 2005. 1995 was taken as a dividing year between two subcorpora, i.e., 1985-1994 *vs*. 1995-2005. To create the corpus, the *BMJ* open-access electronic archive was used.[7] An advanced search by keyword was performed (KW: HIV, sorted by relevance), after which the resulting items were sorted manually on year-by-year basis, in order to extract RAs, i.e., "full-length original research articles, published in the main part of the journal" (Yitzhaki 1997, p. 222), excluding other texts, such as for instance literature review papers. A total of 1250 RATS were collected, 950 of which published in the time span 1985-1994 (subcorpus 1), while 300 in 1995-2005 (subcorpus 2). Table 1 shows the distribution of ATs in the corpus.

Year	**No. items**	**Year**	**No. items**
1985	0	1995	27
1986	34	1996	20
1987	198	1997	28
1988	56	1998	37
1989	135	1999	31
1990	124	2000	21
1991	115	2001	30
1992	157	2002	28
1993	102	2003	34
1994	29	2004	20
		2005	24
Tot. 1985-1994	**950**	**Tot. 1995-2005**	**300**
TOT. 1985-2005			**1250**

Table 1
Distribution of ATs in the corpus.

Assuming that RATs perform key pragmatic functions in terms of informativity/retrievability/attractiveness with respect to the ensuing RA

[5] The HIV and AIDS timelines used in this paper were retrieved from https://www.hiv.gov/hiv-basics/overview/history/hiv-and-aids-timeline.
[6] Materials have been analysed using AntConc (Anthony 2016) and WordSmith Tools (Scott 2017).
[7] Available at http://www.bmj.com/archive. This covers the journal's paper (1840-1994) and online (1995-) archives.

(White, Hernandez 1991; Eyrolle *et al.* 2008), this paper will analyse the strategies enacted by digital, evidence-based medical discourse on HIV.

The analysis will focus on the ways meaning is worded out in conceptual and syntactic terms, and, more specifically, on the way RATs are organised in structural and textual terms. At structural level (Fortanet *et al.* 1998; Haggan 2004; Yitzhaki 1997; Swales 2003; Soler 2007; Jaime Sisó 2009; White, Hernandez 1991; Hjørland, Nielsen 2001), titling constructions will be distinguished into conclusive, interrogative, compound and nominal. By contrasting title construction strategies before and after 1995, the paper will analyse how and to what extent the structural patterning of RATs has been changing in connection with the abovementioned key factors. At textual level, the introduction of expanded nominal phrases in compound titles will be read as a metadiscursive strategy (Hyland 2005; Hartley 2005b, 2007), performing evidential textualization of EBM study design concerns, and thus reflecting changing attitudes towards the production and dissemination of medical knowledge across the 1980s and the 1990s.

3. Results

3.1 No. of RATs/year and AVG sentence length

Table 2 presents an overview of the number of RATs published per year and per subcorpus, as well as the average sentence length per year and per subcorpus.

Year	**No. items**		**AVG s. length**	**Longest**	**Shortest**	**Year**	**No. items**		**AVG s. length**	**Longest**	**Shortest**
1985	0	AVG items/year: 95	--	--	--	1995	27	AVG items/year: 27.7	11.5	24	6
1986	34		8.9	24	3	1996	20		12.9	24	6
1987	198		6.9	24	1	1997	28		14.1	21	10
1988	56		8.4	19	1	1998	37		15.2	28	5
1989	135		8.9	28	2	1999	31		14.6	24	9
1990	124		8.5	21	2	2000	21		14.3	24	6
1991	115		7.8	24	1	2001	30		14.8	39	5
1992	157		8.2	36	1	2002	28		14.5	28	5
1993	102		8.5	24	2	2003	34		14.4	22	7
1994	29		12.19	26	1	2004	20		14.4	28	6
						2005	24		14.7	27	9
Tot. 1985-1994	**950**		**8.7**	**--**	**--**	**Tot. 1995-2005**	**300**		**14.3**	**--**	**--**
Tot. 1995-2005	**1250**								**1.5**	**--**	**--**

Table 2
No. of items/year and AVG sentence length/year.

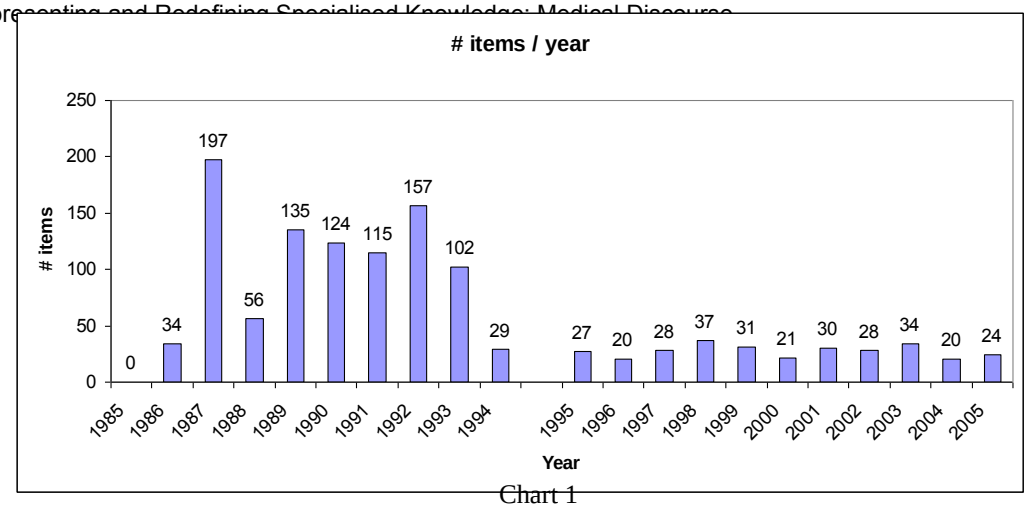


Chart 1
No. items/year.

In 1987, the year the first successful antiretroviral drug (zidovudine AZT) became available, 197 RAs were published; in the 1990s, when AIDS (i.e., the third and final stage of HIV infection) became the object of international epidemiological surveillance, the number of published RAs dropped by almost 70%. Experimental studies on HIV started back in June 1981, when five deaths from an immunodeficiency syndrome, first called "gay cancer" and then GRID, Gay-Related ImmunoDeficiency, were reported in the *Morbidity and Mortality Weekly Report* by the US Centers for Disease Control and Prevention. In 1982, the name AIDS began to circulate in Western medical and media discourse as an aggressive epidemic,[8] progressively involving different population groups, (apparently) starting with male homosexuals, and later involving other categories, such as male and female prostitutes and injecting drug users, and finally involving heterosexual and vertical (mother-child) transmission. The gradual spread of the infection and related pathology is evidenced by titles such as the following (from the first subcorpus):

(1) Willingness of homosexual and bisexual men in London to be screened for human immunodeficiency virus. [1986]
(2) Risk of AIDS related complex and AIDS in homosexual men with persistent HIV antigenaemia. [1987]
(3) Prostitute women and public health. [1988]
(4) Risk behaviours for HIV infection among injecting drug users attending a drug dependency clinic. [1989]
(5) Heterosexually acquired HIV infection. [1989]
(6) Mothers with HIV. [1989]

As shown in Chart 1, the number of published items sharply decreases in 1994, with figures dropping from 102 to 29 the very year AIDS became the

[8] Deaths covered by media speculation include actor Rock Hudson (1985), photographer Robert Mapplethorpe (1989), artist Keith Haring (1990), popstar Freddie Mercury (1991) and dancer Rudolf Nureyev (1993).

leading cause of death in Americans aged 25-44. This may appear as a puzzling circumstance, for which there is no conclusive, univocal explanation. The decrease might be read as a consequence of more advanced knowledge of the virus' behaviour and related pathologies, and/or growing coverage of sensitive areas in social and medical behaviour through the diffusion of guidelines (issued by the US Centers for Disease Control and Prevention) for preventing the diffusion of HIV, and of massive institutional investments in research. As a matter of fact, in 1993 President Clinton established the National Office for AIDS policy at the White House. Also, in June 1994 the FDA approved the first HIV protease inhibitor, which introduced a new era of highly active antiretroviral therapy (HAART). In 1995 saquinavir, a key active ingredient, was approved for prescription use (stage I trials having started in 1989), followed within four months by ritonavir and indinavir, which significantly reduced AIDS death rates within two years – at least in the Western world. We can hypothesize that the introduction of such treatment perspectives might in some way have limited the initial fear of a global AIDS pandemic, although this is mere speculation. What is known for sure is that after 1994, that is, in the second subcorpus, data stabilizes at an average of 27.7 RAs per year.

Trends appear reversed as concerns the average word number per subcorpus, which increases from 8.7 words in 1985-94 to 14.3 words in 1995-2005. Information as to the longest *vs*. shortest constructions is also provided in Table 2, where the shortest constructions between 1985 and 1994 amount to a single word, such as in the following examples:

(7) Casualties. [1987]
(8) Contraception. [1991]

The shortest items in the second subcorpus amount to at least 5 words, while the longest can reach up to 39 words:

(9) Neuropsychiatric complications of nevirapine treatment. [2002]
(10) Prevalence of antibodies to hepatitis B, hepatitis C, and HIV and risk factors in entrants to Irish prisons: a national cross sectional survey: Commentary: efficient research gives direction on prisoners' and the wider public health except in England and Wales. [2001]

As no parameters for title length are to be found in the International Committee of Medical Journal Editors (ICMJE)'s *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals*,[9] or in *BMJ*'s own guidelines for manuscript submission,[10] the

[9] "The title provides a distilled description of the complete article and should include information that, along with the Abstract, will make electronic retrieval of the article sensitive and specific. Reporting guidelines recommend and some journals require that information about the study design be a part of the title

Discussion section of this paper will connect and interpret this data in connection with the two key paradigm shifts taking place at *BMJ* from 1995 on, i.e., the inception of digital communication and of EBM.

3.2 Structural construction of RATs

RATs can be distinguished into four categories, according to different syntactical organizations of the informative material, which can be positioned along a pragmatic continuum between two functions, i.e., efficient information packaging and scientific attractiveness (Sala, Consonni 2018). Table 3 (on the next page) shows the distribution of RATs per year and per subcorpus.

3.2.1 Conclusive titles

Conclusive (full-sentence/declarative) titles are syntactically and semantically autonomous structures, containing finite verbal forms specifying the semantic relationship among the lexical elements in the sentence, as in the following examples:

(11) When things go wrong. [1986]
(12) It is not one of "them"; it is one of all of us. [1988]

In the 1985-94 subcorpus, 17 conclusive titles are present, totalling 1.78%; in 1995-2005, only 2 full-sentence titles can be found (0.67%). This indicates that conclusive titles never appear to have been a popular option for structuring RATs on HIV. Most occurrences in the corpus are, moreover, to be found in the years 1986-88, that is, in the very initial stages of clinical research on the virus. This may be due to the fact that scientific full-sentence titles tend to be related to pragmatic necessities such as informative density/attractiveness, mirroring the researcher's need to quickly inform readers about the contents of the RA, while readers are in turn needing to "know as early as possible in the reading process whether or not the paper contains anything that is of relevance" (Haggan 2004, p. 296). On the other hand, though, conclusive titles may reveal confident assertions, "presented as statement of facts", usually in the present simple tense, reproducing what is known as the "block language" of newspaper headlines (Quirk, Greenbaum 1973); as Table 3 shows, 70.6% of occurrences in the first subcorpus are in the present tense.

(particularly important for randomized trials and systematic reviews and meta-analyses)". Retrieved from http://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html#a.

[10] Available at http://www.bmj.com/about-bmj/resources-authors/forms-policies-and-checklists/title-page.

Year	Conclusive	Interrogative	Nominal	Compound
1985	0	0	0	0
1986	4 *pres. tense: 3*	0	27	3
1987	6 *pres. tense: 4*	4	132	55
1988	3 *pres. tense: 1*	1	36	16
1989	1 *pres. tense: 1*	2	106	26
1990	1 *pres. tense: 1*	2	92	29
1991	2 *pres. tense: 2*	7	71	35
1992	0	4	115	38
1993	0	8	67	27 *EBM in exp. NP: 4*
1994	0	1	23	5
Tot. 1985-1994	**17 (1.78%)** ***pres. tense: 12 (70.6%)***	**29 (3.05%)**	**669 (70.42%)**	**234 (24.64%)** ***EBM in exp. NP: 4 (1.7%)***
1995	0	1	15	12 *EBM in exp. NP: 3*
1996	0	1	10	9 *EBM in exp. NP: 4*
1997	0	0	7	21 *EBM in exp. NP: 8*
1998	0	0	10	27 *EBM in exp. NP: 19*
1999	0	0	8	23 *EBM in exp. NP: 14*
2000	0	0	5	16 *EBM in exp. NP: 14*
2001	0	0	4	25 *EBM in exp. NP: 18*
2002	0	1	7	20 *EBM in exp. NP: 19*
2003	0	0	5	29 *EBM in exp. NP: 27*
2004	2 *pres. tense: 2*	0	1	17 *EBM in exp. NP: 15*
2005	0	1	6	17 *EBM in exp. NP: 17*
Total 1995-2005	**2 (0.67%)** ***pres. tense: 2 (100%)***	**4 (1.33%)**	**78 (26%)**	**216 (72%)** ***EBM in exp. NP: 158 (73.15%)***

Table 3
Distribution of structural constructions/year/subcorpus.

This may indicate "confident optimism projected by the writer that what he is reporting stands true for all time or is not simply a one-off occurrence", as though the researchers were conveying "the certainty that the method, measurements, calculation etc. employed have yielded impregnable findings" (Haggan 2004, p. 297). Occurrence of conclusive titles in the 1995-2005 subcorpus is in fact accompanied by the use of hedges, especially in the form of the modal verb *may*, which limits the scientist's claim for credibility, as in the following example:

(13) Acquired haemophilia A *may* be associated with clopidogrel. [2004; emphasis added]

3.2.2. Interrogative titles

Interrogative titles are formulations constructed as questions, conveying meanings interrogatively rather than assertively, thus either pointing out possible cognitive gaps to be dealt with in the ensuing RA, which the reader might wonder about, or casting doubts over previous research conclusions. In this respect, interrogative titles typically express "queries in need of reply, interpretation, and conclusion" (Soler 2007, p. 100), as in the following examples:

(14) After safe sex, safe surgery? [1987]
(15) How informed is patients' consent to release of medical information to insurance companies? [1989]
(16) Is risk of Kaposi's sarcoma in AIDS patients in Britain increased if sexual partners came from United States or Africa? [1991]

Since interrogative RATs may be considered as syntactical expressions of doubt, paralleling in some way medical research as a question process, it seems coherent that they represent only 3.05% of the 1985-94 subcorpus (29 occurrences), dropping to 1.33% in the second subcorpus (4 occurrences) and remaining nearly silent after 1997.

3.2.3 Nominal titles

Nominal titles are structures either consisting of single verbless expressions, or containing non-finite verbal forms (such as gerund, participle, *to* + infinite, etc.). These are typical of "block language" (Straumann 1935), '*headlinese*' (Garst, Berstein 1963), or economy grammar (Halliday 1967), and often found in contexts with fixed space constraints – such as advertising, book titles, and newspaper headlines. They are generally associated with the omission of auxiliaries (*be*, *have*, *do*) and articles (*a/an*, *the*), and a preference for passive voice and nominalization, as can be observed in the following examples, taken from both subcorpora:

(17) AIDS, them, and us. [1987]
(18) Female streetworking prostitution and HIV infection in Glasgow. [1992]
(19) Prevalence of HIV and injecting drug use in men entering Liverpool prison. [1998]
(20) Cost effectiveness analysis of strategies for maternal and neonatal health in developing countries. [2005]

In the 1985-94 subcorpus, nominal constructions are dominant, represented by 669 items (70.42%) and followed by compound titles (24.64%), whereas proportions become inverted in the 1995-2005 subcorpus, where nominal titles drop to 26% (78 out of 300 occurrences) and compound titles increase to 72% (216 items). As new discoveries and advancements were being made in HIV research, as it were, nominal syntax probably no longer seemed to be the most appropriate strategy, for it is clear from Table 3 that nominal titles become recessive in the 1995-2005 subcorpus, to the benefit of compound constructions.

3.2.4 Compound titles

Compound (colonic/hanging, Hartley 2005b) titles are composed of two semantically related parts (phrases, clauses or full sentences, both declarative and interrogative) typically joined by a colon, full stop, dash or other punctuation mark (Hartley 2007, p. 553). In terms of thematic structure, they are organized as theme-rheme clusters, where the former part of the title introduces the RA's topic and the latter one – usually an expanded noun phrase, in which particular aspects of the topic to be dealt with are specified – highlights its relevance by framing it in 'general-specific', 'cause-effect', 'problem-solution', 'research question-research method' patterns. Instances of compound titles from both subcorpora are provided below:

(21) Campaign against AIDS in Switzerland: evaluation of a nationwide educational programme. [1986]
(22) Infertility management in HIV positive couples: a dilemma. [1991]
(23) Risk of HIV related Kaposi's sarcoma and non-Hodgkin's lymphoma with potent antiretroviral therapy: prospective cohort study. [1999]
(24) Treatment exhaustion of highly active antiretroviral therapy (HAART) among individuals infected with HIV in the United Kingdom: multicentre cohort study. [2005]

As already mentioned, while the majority of RATs in the former subcorpus are nominal in structure, the trend is reversed from 1995 on: Table 3 shows that in 1995-96 the proportion is more evenly balanced, with nominal titles still slightly outnumbering compound titles (15 and 10 *vs*. 12 and 9 respectively), but as of 1997 figures steadily confirm the predominance of compound over nominal structures. In 2004 only one nominal title was published *vs*. 17 compound titles. As will be argued in the Discussion section of this paper, the increasing preference for compound syntax in the later

subcorpus may again be related to the communicative and epistemological shift brought about in the mid-1990s by electronic literacy and the EBM paradigm.

3.3 Information patterning in compound titles

Table 3 also shows a significant change in the strategies that compound titles tend to use in order to package/sequence information for readers. Such

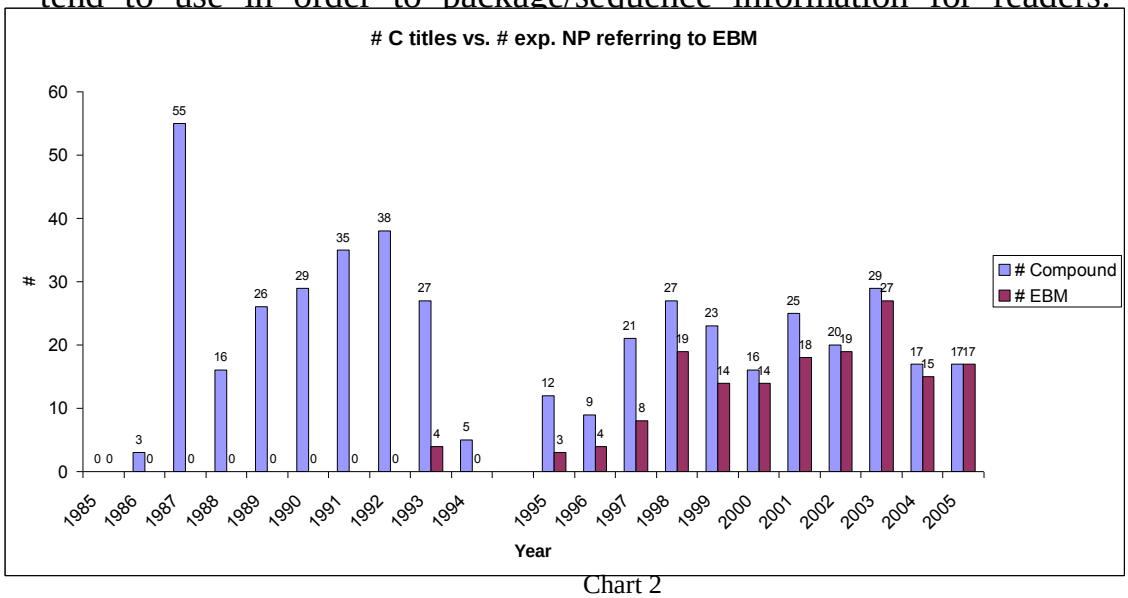


Chart 2
No. of compound titles *vs.* No. of expanded noun phrases focusing on EBM study design.

Provided that the thematic part of compound titles generally focusses on the clinical topic to be dealt with in the RA, in the former subcorpus the expanded noun phrase following the colon (and occupying the rheme/filler position) covers a range of topics, eliciting the reader's curiosity, which mainly concern HIV or its development into AIDS. These may range from details about the infection's onset, progress and geography, to social groups involved in the epidemic, to specific variables linked to clinical aspects of the disease; but nothing in compound titles in the years 1985-94 seems to specifically refer to the methodology of research employed in the ensuing RA. The most frequent topics seem generally related to epidemic details or pathways to possible treatment, as in the examples below:

(25) AIDS: a faltering step. [1987]
(26) Surveillance of AIDS cases: how acceptable are the figures? [1988]
(27) Early HIV infection: to treat or not to treat? [1990]
(28) No escape: HIV transmission in jail. [1993]

Conversely, the 1995-2005 subcorpus shows an increasing number of rhematic noun phrases explicitly referring to EBM practice and study design features, i.e., the methodology following which the research was conducted,

which proves a crucial factor in a RA's critical appraisal, that is, its hierarchical evaluation in terms of clinical evidence and scientific prestige. In such noun phrases, specific reference is made to EBM study design within the hierarchy of evidence, which the reader is invited to check out and assess by reading the Method section. In the years 1995-97, approximately 30% of rhematic noun phrases focus on study design terminology, as in the following examples:

(29) Does the onset of tubercolosis in AIDS predict shorter survival? Results of a cohort study in 17 European countries over 13 years. [1995]
(30) Mortality associated with HIV-1 infection over five years in a rural Ugandan population: cohort study. [1997]

The percentage rapidly grows to around 60% of occurrences in 1998-99, while from 2000 on nearly 100% of compound titles refer to EBM study design, which tends to occupy the whole filler slot at the expense of previously foregrounded details (e.g. geographical or social variables involved in the research). That is to say, in the later subcorpus the rhematic/new information part of compound titles no longer focuses on HIV infection *per se*, but on global HIV control through massive evidence-based research and therapy, as in the following examples:

(31) Effect of zinc supplementation on malaria and other causes of morbidity in West African children: randomised double blind placebo controlled trial. [2001]
(32) Effect of iron supplementation on incidence of infectious illness in children: systematic review. [2002]
(33) Stable partnership and progression to AIDS or death in HIV infected patients receiving highly active antiretroviral therapy: Swiss HIV cohort study. [2004]
(34) Treatment exhaustion of highly active antiretroviral therapy (HAART) among individuals infected with HIV in the United Kingdom: multicentre cohort study. [2005]

In these structures, the sequential "add-on" theme/rheme patterning indicates the positioning of each piece of research – such as, for instance, a cohort study, RCT, systematic review, etc. – within the EBM paradigm, and tends to coincide with the structure's textualization in terms of Information Unit. The thematic part of the title (the given part of the message) usually refers to a specific clinical aspect of HIV. Interestingly, very few titles still focus on the aetiology of the virus after 1995, as this had probably been clarified by previous research, while most deal with prolonging life expectancy through combined antiretroviral treatment, and/or with the neutralization of AIDS's most aggressive consequences, especially in developing countries. The rhematic part (the new part of the message) more and more tends, on the other hand, to conspicuously coincide with the research's study design.

4. Discussion

The phenomena identified and quantified so far can be discussed in relation to the two key factors considered in the research questions of this paper, that is, the impact of the Internet and digital literacy, and of EBM clinical protocols, on the codification and transmission of written medical discourse about HIV.

As concerns the average title length (cf. Section 3.1, Table 2 and Chart 1 above), both factors can be evidenced as influencing the patterns and variations of RATs between the subcorpora. With respect to *BMJ*'s migration from paper to server, the brevity of titles prior to 1995 may be due to the constraints of limited space in the printed edition of the journal, with "the resulting need to be brief and succinct" (Haggan 2004, p. 294). On the contrary, increasing length in the second subcorpus may indicate a steady growth in RAT's informative content, compatible with increased space availability in online publication (which would agree with results presented in Berkenkotter, Huckin 1995, and mirror a common "time factor" trend in scientific titles, as evidenced in Yitzhaki 1997, p. 221). Finally, and importantly, the length of a title is crucial to its online retrieval; the longer the title, the more lexical items it contains, and the greater the chances that it may be retrieved by a query.

Alongside the changes brought about by digital publication, the data may also be explained following the evolution in HIV research and knowledge during the 1990s. As a field of research becomes more complex, RATs are actually expected to become longer and to mirror "the development, refinement, and extension both of underlying theories and of more and more complex research methods and procedures" (White, Hernandez 1991, p. 731). As evidenced in Hjørland, Nielsen (2001, p. 266), although the hard sciences traditionally tend to have longer, more informative titles than softer and popular sciences, the increase in average sentence length observed in the present corpus may be due to "increasing specialization in research, creating a need for more words to express a given piece of research" (*ibid.*). This seems compatible with the onset of EBM at *BMJ* from 1995 on, as longer and more complex titles function as vehicles to disseminate a whole new medical epistemology.

Concerning the patterns and variations evidenced among the four syntactical categories of RATs in the corpus, the impact of digital literacy and EBM can be observed at different levels. The different frequency patterns of conclusive titles between the subcorpora (see Section 3.2.1 above) may firstly suggest a conflation in RATs between scientific and promotional language, especially where '*headlinese*' effects are employed to express some degree of epistemological certainty on the topic. In the case of HIV research,

conclusive titles may be hypothesized to mirror the assertive/urgent tone of initial research, that is, in the former subcorpus, when scientific interest was mainly concerned with the transmission of the virus (initially involving certain stigmatized social categories), and before the actual complexity of multiple aetiological and clinical factors was taken into serious consideration. This can be confirmed by the fact that the use of conclusive sentences seems to disappear in the corpus as of 1991. The same trend is furthermore shown by the frequency of interrogative titles (cf. Section 3.2.2 above), which seems to confirm the results in Soler (2007, p. 100), and to reflect lesser need for the structural expression of scientific dilemma as time went by, from the mid-1980s to the late 1990s, when more decisive research on the virus was being carried out and the paradigm shift from traditional practice to EBM was well on its way.

The opposite incidence of nominal structures in the subcorpora (cf. Section 3.2.3 above) may in turn be interpreted as linked to both factors taken into consideration in this paper. The frequency of nominal constructions in the first subcorpus, with their high capacity for showcasing a discipline's substantial keywords, may be traced to the scientific need for lexicalization strategies in the early years of research, when HIV became an increasingly delicate social topic, as more research was being carried out, showing more complex aetiological factors and more detailed hypotheses concerning the progress of AIDS. The high prevalence of nominal structures may in this respect be associated to the prototypical classificatory nature of medical science, which tends to treat its object of study in taxonomical fashion (Soler 2007, p. 101). This seems to be a result shared by Haggan (2004, p. 307), who concludes that a noun phrase, accompanied by one or more post-modifying prepositional phrases and/or moderate to heavy pre-modification, is the most popular choice for traditional scientific title-patterning, guaranteeing that RATs attain both informative precision/explicitness (provided by the piling up of post-modifiers) and block-language-effect attractiveness (provided by shorter and generally more evenly balanced pre-modified structures; see also Rush 1998).

On the other hand, though, the increasing incidence of compound syntax from 1995 on (as shown in Section 3.2.4 and Chart 2 above) seems to mirror the impact of the new literacy standard brought about by digital communication in the mid- and late 1990s, whereby the use of the Internet as the main channel for knowledge articulation and dissemination has triggered significant changes in highly specialized discourse, from markedly standardised, crystallised formulations – meant for information filtering before lay dissemination – to more articulated ones, placing emphasis on distinctively argumentative, persuasive and metadiscoursive functions. Traditional informativity is thus complemented by attractiveness, which may

suggest further research into EBM communication as an interdiscursive area between scientific and advertising language (Haggan 2004; Hartley 2007; Bhatia 2004), thus paralleling and enriching the potential hybridity traditionally inherent to the use of conclusive – or '*headlinese*' – medical RATs (cf. Section 3.2.1 above).

Moreover, compound titles contain an increased number of lexical items, which on the one hand may be useful to retrieve RAs in online searches and specialized databases, while, on the other, providing room for showcasing essential research advancements, thus contributing to the diffusion of new knowledge and to its electronic retrieval. Whereas paper RATs are usually printed on the same page as, or in the vicinity of, the full RA, so that the correspondence between the research piece and its title is immediately clear, online textuality separates the title from the article, which is usually on a different webpage, for which reason the title needs to become at once a more informative (i.e., longer) and more autonomous structure. No longer ancillary to the ensuing RA, a compound title is in itself a semantically full textual typology, activating specific processing dynamics which can facilitate the decoding of the RA, including "attentional focusing during reading", "encoding of the text structure", governing "text summary and recall", determining "the relative importance of information supplied in a text", integrating "text information by establishing relations between different elements", and contributing "to the building of [readers'] cognitive representation" (Eyrolle *et al.* 2008, p. 242).

As noted in Hartley (2007, p. 558), compound titles allow writers to both attract and inform readers: this is achieved by means of the theme/rheme (or gap/filler) information sequencing they provide, whereby the reader's curiosity is engaged by the thematic part of the cluster (presenting a research question) and the filler slot is occupied by the rhematic part (offering insight into how the question will be addressed in the RA). The first part indicates the research area covered by the RA, while the second narrows down on the research's specifics, especially as concerns clinical applications of the topic, or other details concerning its positioning within the discipline (Haggan 2004, p. 302). In opposition to the traditional nominal structure – where findings are presented synoptically (usually through heavy pre-modification or the piling up of prepositional post-modifiers, which provide a mapping of the topic and findings) – compound titles follow a sequential "add-on" theme/rheme patterning, pivoting on the opposite principle, i.e., the principle of "presumption of ignorance" (*ibid.*). The writer must first present a hypothesis regarding his readers' knowledge of the topic/field of research, after which he has to draw their attention towards what he presumes they are ignorant/in need of, following the shortest path to easing the reader's processing of the text.

This represents an efficient system for both information packaging and attention drawing, which marks a dramatic change in the pragmatic purposes of expert-to-expert communication, from the elitist, gate-keeping, peer-to-peer traditional exchange of clinical practice (potentially viewed as bias after the inception of EBM) to the sharing of the best available evidence, where personal experience and bench studies rank low in the hierarchy of evidence. By performing both informative and attractive functions, as well as by revealing knowledge dissemination as a negotiation between hypotheses and expectations, compound titles can be read as a marker of a scientist/writer's own self-aware, negotiated positioning with respect to both Web literacy and the EBM hierarchy of evidence.

This trend seems to be confirmed by an increasingly frequent textualization strategy shown by RATs in the late 1990s, i.e., the packaging of methodological information in the rhematic part of the cluster (cf. Section 3.3 above). Such textualization strategy may be said to appear in the 1995-2005 subcorpus as a consequence of EBM implementation, and can be read on a metadiscursive level as a marker of evidentiality, i.e., a textual strategy signalling "the source of speaker's knowledge" (Johnstone 2009, p. 30) through "the ascription of information or opinion in a text to sources which may be animate or inanimate", such as a piece of empirical research, a clinical trial or a laboratory experiment (Hunston 2003, p. 181). By framing RATs within the EBM hierarchy of evidence, the expanded rhematic noun phrase in compound titles from 1995 to 2005 functions as a marker of discourse legitimization in the context of the new epistemic paradigm brought about by the inception of EBM.[11] Conversely, the general directional/geographic/social details provided in compound titles before 1995 (with the exception of the four nominal phrases conveying EBM practices in 1993) may, after the mid-1990s, appear as tokens of pre-EBM "bias", therefore progressively becoming recessive textualization resources.

5. Concluding remarks

This paper has aimed to identify and quantify the key syntactical and textual features of RATs dealing with HIV, with reference to the epistemological paradigm brought about in the mid-1990s by the onset of Evidence-Based Medicine, and to the digital literacy standard established by the use of the Internet as the main channel for contemporary knowledge dissemination. The

[11] This seems confirmed by the introduction of a rule in the preparation of new manuscripts for *BMJ*, according to which all research papers should include a description of its study design. Retrieved from http://www.bmj.com/about-bmj/resources-authors/forms-policies-and-checklists/title-page.

traditionally codified discourse of clinical pathology in highly specialized contexts such as the *BMJ* seems to have undergone major changes from the mid-1990s on, as significant trends have been highlighted by contrastive analysis between the two subcorpora: decreasing number of RAs published on HIV (950 *vs*. 300); increasing sentence length (8.7 *vs*. 14.3); opposite incidence of nominal and compound syntactical structures (respectively 70.42% *vs*. 24.64% in 1985-1994 and 26% *vs*. 72% in 1995-2005); increasing occurrence of EBM-related rhematic noun phrases in compound titles in the later subcorpus, from 1995-97 (30%) to 1998-99 (60%) to 2000 and beyond (100% in 2005).

The present data seems to suggest that major changes have been occurring in the process of knowledge dissemination within specialized discourse in the last thirty years, due to both factors taken in consideration in this paper. On the one hand, medical communication has found a strategic genre in RATs, which have become an increasingly impactful resource/convention for the sharing of clinical information meant for expert users. In particular, beside performing a key pragmatic function with respect to informativity and attractiveness, especially in the digital environment, the diffusion of compound titles provides an instant description of both the clinical topic addressed in the RA (to be identified with the thematic/given part of the structure's thematic sequence) and the study design employed to investigate it (to be found in the rhematic/new part of the structure). At the same time, compound titles allow readers and fellow researchers to rank the evidence provided in the RA within the EBM hierarchy. This means that, even before reading the actual abstract to the paper, readers can form an idea of what it will be about and what impact its results can be expected to have in terms of methodological credibility. Beside the traditional pragmatic functions of informativity and attractiveness, RATs thus seem to have increasingly developed a third and crucial function: an epistemological one.

Being a pilot study, this paper has compiled and analysed a corpus of titles from one source only (albeit an authoritative one). It is clear, however, that further research in medical linguistics related to the clinical and cultural history of HIV would benefit from the use of larger and more heterogeneous corpora. These may include journals from different cultural milieus such as, for instance, Europe *vs*. the USA, as well as from different scientific perspectives and epistemological coordinates, sampling publications with, for instance, different institutional affiliations and Impact Factors, etc. The use of larger and more comprehensive and articulated corpora would allow to look further into the linguistic and representative dissemination of HIV from a wider – and more interdisciplinary – angle.

The present data seems, however, to indicate that the onset of new scientific and literacy paradigms in the mid-1990s has progressively required

medical expository practices to finetune their communicative skills, and in particular to showcase as much information as possible as regards the methodological design of each piece of research that is published in expert-to-expert contexts such as the *BMJ*. By simply browsing digital search results, and by simply reading a compound title, qualified readers and fellow researchers will immediately know where to rank a piece of research into the hierarchy of evidential knowledge. RATs therefore seem to pragmatically activate scientifically effective expectation protocols in a specialized audience.

Bionote: Stefania Consonni is a researcher in English Language at the University of Bergamo. Her publications focus on textual paradigms, narratology, the semiotics of visual *vs*. verbal language, specialized communication in a discourse-analytical perspective, and the pragmatic features of traditional and digital genres within academic and professional discourse. She is an active member of the Research Centre on Languages for Specific Purposes (CERLIS), based at the University of Bergamo, and has been involved in inter-academic Research Projects, funded by the Italian Ministry of Education, on social and cultural studies and on specialized discourse.

Author's address: stefania.consonni@unibg.it

Acknowledgments: I would like to thank Dr. Francesco Testa for providing many insightful remarks, and for lending his clinical expertise to the drafting of this paper. Any mistakes and/or inaccuracies are my own.

References

Anthony L. 2016, *AntConc* (Version 3.4.4) [Computer Software], Waseda University, Tokyo.

Berkenkotter C. and Huckin T.N. 1995, *Genre Knowledge in Disciplinary Communication: Cognition/Culture/Power*, Lawrence Erlbaum, Hillsdale.

Bhatia V.K. 2004, *Worlds of Written Discourse: A Genre-Based View*, Continuum, London.

Bucchi M. 1998, *Science and the Media: Alternative Routes in Scientific Communication*, Routledge, New York.

Calsamiglia H. 2003, *Popularization Discourse*, in "Discourse Studies" 5 [2], pp. 139-146.

Calsamiglia H. and Van Dijk T.A. 2004, *Popularization Discourse and Knowledge about the Genome*, in "Discourse & Society" 15 [4], pp. 369-389.

Eyrolle H., Virbel J. and Lemarié J. 2008, *Impact of Incomplete Correspondence between Document Titles and Texts on User's Representations: A Cognitive and Linguistic Analysis Based on 25 Technical Documents*, in "Applied Ergonomics" 39, pp. 241-246.

Fortanet I., Posteguillo S., Coll J.F. and Palmer, J.C. 1998, *Linguistic Analysis of Research Article Titles: Disciplinary Variations*, in Vázquez I. and Guillén I. (eds.), *Perspectivas Praguietices en Linguistica Aplicada*, Anubar, Zaragoza, pp. 443-447.

Foucalt M. 1963, *Naissance de la clinique. Une archéologie du regard médical*, P.U.F., Paris.

Garst R. and Berstein T. 1963, *Headlines and Deadlines*, Columbia University Press, New York.

Garzone G. 2006. *Perspectives on ESP and Popularization*, CUEM, Milano.

Genette G. 1988, *Structure and Functions of the Title in Literature*, in "Critical Inquiry" 14 [4], pp. 692-720.

Giannoni D.S. 2014, *Whose Genre Awareness? The Case of Medical Titles*, in O'Rourke B., Bermingham N. and Brennan S. (eds.), *Opening New Lines of Communication in Applied Linguistics*, Scitsiugnil Press, London, pp. 151-160.

Gotti M. 2003, *Specialized Discourse*, Peter Lang, Bern.

Gotti M. 2013, *The Analysis of Popularization Discourse: Conceptual Changes and Methodological Evolutions*, in Kermas S. and Christiansen T. (eds.), *The Popularization of Specialized Discourse and Knowledge across Communities and Cultures*, Edipuglia, Bari, pp. 9-32.

Gotti M., Maci S.M. and Sala M. (eds.) 2015, *Insights into Medical Communication*, Peter Lang, Bern.

Greenhalgh T. 2010, *How to Read a Paper: The Basics of Evidence-Based Medicine*, Wiley/Blackwell, Oxford.

Haggan M. 2004, *Research Paper Titles in Literature, Linguistics and Science: Dimensions of Attraction*, in "Journal of Pragmatics" 36, pp. 293-317.

Halliday M.A.K. 1967, *Intonation and Grammar in British English*, Mouton, The Hague.

Hartley J. 2005a, *Improving that Title: The Effects of Colons*, in "European Science Editing" 31, pp. 45-47.

Hartley J. 2005b, *To Attract or to Inform: What Are Titles for?*, in "Journal of Technical Writing and Communication" 35 [2], pp. 203-213.

Hartley J. 2007, *Planning that Title: Practices and Preferences for Titles with Colons in Academic Articles*, in "Library & Information Science Research" 29, pp. 553-568.

Hunston S. 2003, *Evaluation and the Planes of Discourse: Status and Value in Persuasive Texts*, in Hunston S. and Thompson G. (eds.), *Evalutation in Texts: Authorial Stance and the Construction of Discourse*, Oxford University Press, Oxford, pp. 176-207.

Hyland K. 2005, *Metadiscourse: Exploring Interaction in Writing*, Continuum, London.

Hjørland B. and Nielsen L.K. 2001, *Subject Access Points in Electronic Retrieval*, in "Annual Review of Information Science and Technology" 35, pp. 249-298.

Jaime Sisó M. 2009, *Anticipating Conclusions in Biomedical Research Article Titles as a Persuasive Journalistic Strategy to Attract Busy Readers*, in "Miscelánea" 39, pp. 29-51.

Johnstone B. 2009, *Stance, Style, and the Linguistic Individual*, in Jaffe A. (ed.), *Stance: Sociolonguistic Perspectives*, Oxford University Press, Oxford, pp. 29-62.

Kozminsky E. 1977, *Altering Comprehension: The Effect of Biasing Titles on Text Comprehension*, in "Journal of Applied Research in Memory and Cognition" 5 [4], pp. 482-490.

Martin S. 1998, *How News Gets from Paper to Its Online Counterpart*, in "Newspaper Research Journal" 19 [2], pp. 64-73.

Myers G. 2003, *Discourse Studies of Scientific Popularization: Questioning the Boundaries*, in "Discourse Studies" 3, pp. 265-279.

Quirk R. and Greenbaum S. 1973, *A University Grammar of English*, Longman, London.

Raffo M. 2016, *Translation and Popularization: Medical Research in the Communicative Continuum*, in "Meta" 61, pp. 163-175.

Rush S. 1998, *The Noun Phrase in Advertising English*, in "Journal of Pragmatics" 29, pp. 155-171.

Sackett D.L., Rosenberg W.M.C., Gray J.A.M., Haynes R.B. and Richardson W.S. 1996, *Evidence Based Medicine: What It Is and What It Isn't*, in "British Medical Journal" 312, pp. 71-72.

Sala M. and Consonni S. 2018, *Article Titles in Online Medical Popularization*, in Bondi M., Cacchiani S. and Cavalieri S. (eds.), *Knowledge Dissemination at a Crossroads: Genres and New Media Today*, Cambridge Scholars, Newcastle upon Tyne, pp. 1-20.

Scott M. 2017, *WordSmith Tools version 7*, Lexical Analysis Software, Stroud.

Shinn T. and Whitley R. (eds.) 1985, *Expository Science: Forms and Functions of Popularisation*, Reidel, Dordrecht.

Smith R. 2000, *Informative Titles in the BMJ*, in "British Medical Journal" 320, p. 915.

Soler V. 2007, *Writing Titles in Science: An Exploratory Study*, in "English for Specific Purposes" 26, pp. 90-102.

Straumann H. 1935, *Newspaper Headlines: A Study in Linguistic Method*, Allen & Unwin, London.

Swales J. 2003, *Genre Analysis: English in Academic and Research Settings*, Shanghai Foreign Language Education Press, Shanghai.

Yitzhaki M. 1997, *Variations in Informativity of Titles of Research Papers in Selected Humanities Journals: Comparative Study*, in "Scientometrics" 38 [2], pp. 219-229.

Wang Y. and Bai Y. 2007, *A Corpus-Based Syntactic Study of Medical Research Article Titles*, in "System" 35, pp. 388-399.

White A. and Hernandez N.R. 1991, *Increasing Field Complexity Revealed through Article Title Analyses*, in "Journal of the American Society for Information Science" 42 [10], pp. 731-734.

Zeiger M. 1991, *Essentials of Writing Biomedical Research Papers*, McGraw-Hill, New York.

DOES MEAT CAUSE CANCER?
The discursive construction of meat carcinogenicity in a corpus of scientific texts

SABRINA FUSARI

Abstract – In 2015, the International Agency for Research on Cancer (IARC) published a report on the carcinogenicity of red and processed meat, incorporating red meat in Group 2A carcinogens (probably carcinogenic to humans) and processed meat in Group 1 (carcinogenic to humans). This announcement attracted immediate interest from other scientists, especially in medical research, where the relation between cancer and food has been investigated extensively for many years. This paper aims to analyze the discursive construction of meat carcinogenicity in a set of scientific papers published in the wake of the IARC communiqué. For this purpose, an electronic corpus was assembled from a range of academic journals featured in the database *Elsevier Science Direct*, for a total of 384,491 words, which were fully POS-tagged, partially parsed using a systemic functional grammatical formalism, and subsequently analyzed on Antconc. The methodology adopted to analyze these data is a combined corpus assisted discourse analysis approach, focusing mainly on experiential noun group structures, specifically those involved in patterns of nominalization, which typically aim to achieve monorefentiality in scientific discourse. However, in this corpus, the denotational boundaries of *meat* (what animal-based foods count as *meat* or *meat products*; what animals have *red* rather than *dark* or *white* meat; the exact nature of *meat processing*) are not entirely clear, and this "semantic debate" (Lippi *et al.* 2016, p. 2) is central to the preoccupations of medical and nutrition experts. Therefore, conclusions show that linguists could make a useful contribution to cancer science by devising a set of universally agreed definitions of meat types, so as to agree on the level of health risk that each may cause.

Keywords: corpus-assisted discourse analysis; Systemic Functional Linguistics; cancer; meat; IARC.

1. Introduction

This paper presents a corpus-based study of the discursive reaction of the scientific community[1] to the publication, in October 2015, of a report entitled

[1] As detailed in Section 3 of this paper, we consider agricultural, biological, biochemical, genetic, environmental, medical, dental and nursing sciences, as defined by categories used in the *Elsevier Science*

Doi Code: 10.1285/i9788883051531p71

Carcinogenicity of consumption of red and processed meat, issued by the International Agency for Research on Cancer, the IARC (International Association for Research on Cancer).

The aim of this study is to investigate how meat carcinogenicity is not only described in scientific papers published in the wake of this IARC communiqué, but also constructed, both scientifically and discursively.[2] The research questions we address involve the identification of the main discursive features used to construct meat carcinogenicity, and especially the reference of *meat* and its byproducts in extralinguistic reality. This concern is shared by linguistic and medical studies, as both undertake to achieve a univocal categorization of animals that provide *red meat*, and an unambiguous definition of *meat processing*. As a matter of fact, what counts as *red meat* and *meat processing* is not an objective datum, either in discourse or in science, as the reference of these expressions may vary across different languages and cultures.

To achieve this goal, firstly, we provide some theoretical background to this study, both sociocultural (the role, extent and understanding of meat eating in human nutrition today, according to a number of academic and popular scientific sources) and linguistic (ecolinguistic approaches to the discursive construction of animals as food). Secondly, we describe the dataset that was assembled for this study, and the way it was tagged and explored on corpus programs, to bring out significant patterns in the discursive construction of meat as a potential carcinogen. Thirdly, we illustrate the main findings of our study, focusing on the structure of the noun group, which exhibits a series of features of scientific language, typically aimed to achieve monorefentiality in the choice of terminology. Indeed, our data suggest a denotational uncertainty as to what exactly qualifies as *meat*, i.e.

- whether it includes poultry and fish;
- what animals correspond to the various colours of meat, i.e. whether pork should be considered *red*, and therefore identified as a potential carcinogen by the IARC, or if younger animals are better classified as *white*, and therefore uninvolved in this issue;
- finally, what exactly is meant by *processing*, i.e. whether *processed meat*, which the IARC considers certain to increase the risk of cancer, only

Direct database. Although the majority of the studies are medical, the sample includes all the articles which referenced the IARC report on meat carcinogenicity in or before August 2017.

[2] In this study, the notion of 'discursive construction' refers generally to the basic pragmatic concept that "speaking is doing", traceable to Austin's observation that "saying something will often, or even normally, produce certain consequential effects upon the feelings, thoughts, or actions of the audience" (Austin 1962, p. 101). Therefore, although the author is aware that this framework has been widely used in critical and socioconstructivist discourse theory, these further theoretical elaborations of the role of speech acts in constructing reality are beyond the scope of this paper.

includes foods made from pork, like bacon and salami, or also poultry-based ones, like chicken sausages and cold cuts obtained from turkey or fowls.

Finally, we present our conclusions, by addressing a recurrent issue in the scientific articles analyzed, i.e. the authors' surprisingly high level of metalinguistic awareness of the semantic conflict around what counts as meat, and also around what should be considered as evidence that meat may represent a health hazard. One article in our sample (Lippi *et al.* 2016) explicitly calls for cooperation between clinical scientists and linguists towards the definition of a set of universally agreed definitions of meat, trying to overcome the problem of culture-specificity in the understanding of what animals are good to eat, and/ or correspond to various meat colours. Suggestions are therefore made to take up this challenge, so as to enable consumers to receive more objective information than they can access now about the level of health risk that each animal-based food may cause.

2. Theoretical background

A recent *National Geographic Education* project, entitled *What the World Eats*,[3] has highlighted a generalized increase in the consumption of meat worldwide over the past few decades, especially in countries which have joined the capitalist society only in recent times, like China, but also in parts of the world that have a traditionally meat-rich diet. For example, in the United States, global meat consumption per person has increased by 30% between 1961 and 2011, despite public awareness of the risks of cholesterol, saturated fats, and other nutrients that especially certain meats are rich of: these health scares have apparently not affected the American consumer's hunger for meat, as beef alone has grown by 50% in terms of tons consumed over the five decades considered. Although, interestingly, seafood is considered a type of meat[4] in this *National Geographic* project, no attempt is made to categorize meat types according to colour: this, instead, is a fundamental preoccupation of the IARC study under discussion, as its findings state that *red* meat is likely to increase cancer risk, but they do not mention white. Figure 1 below, taken from a *Cancer Research UK* (CRUK)

[3] The project illustrates a detailed breakdown of food types and nutrients eaten by people in various countries of the world, in terms of grams and calories, with a special section about meat consumption, divided into types of animals eaten. The project, built in conjunction with the *National Geographic* series *Future of Food*, and based on FAO statistics, is freely accessible at https://www.nationalgeographic.com/what-the-world-eats/ .

[4] Fish is known to have an "ambiguous position" (Montanari 2015, p. 72) in many food cultures, probably because it is a Christian symbol. In fact, in the Middle Ages, eating fish was admissible during Lent and other 'lean' periods of the year, while dairy and eggs were excluded, due to their being excrete by animals.

commentary of the IARC report, and based on the IARC classification of carcinogens, illustrates a possible association between animals, meat types and processed food items.

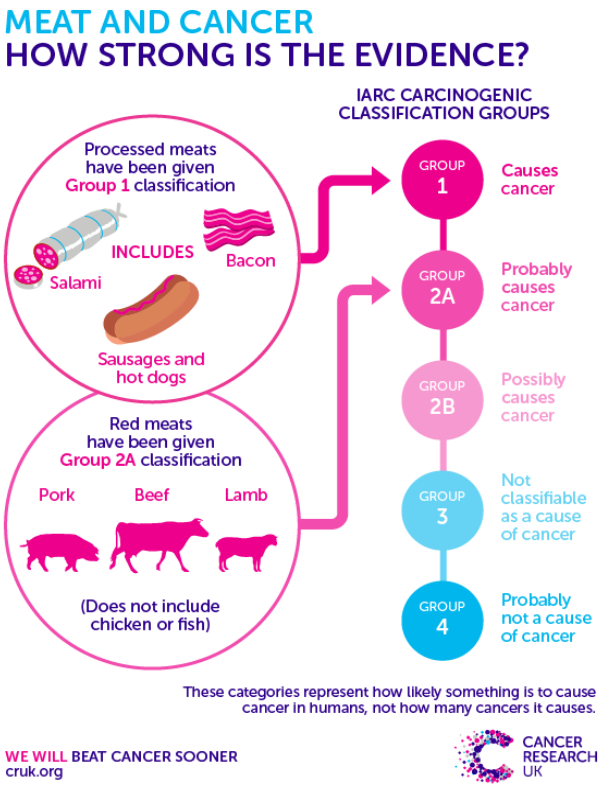


Figure 1
CRUK infographic on the relation between meat and cancer (Dunlop 2015).

Figure 1 suggests quite a neat and tidy breakdown of meat types in precise correspondence with animal species, with pork, beef and lamb qualifying as *red*, and chicken and fish being explicitly excluded from the IARC classification. However, this schematization is at least partially at odds with what the IARC itself writes in its report.

> Red meat refers to unprocessed mammalian muscle meat – for example, beef, veal, pork, lamb, mutton, horse, or goat meat – including minced or frozen meat; it is usually consumed cooked. Processed meat refers to meat that has been transformed through salting, curing, fermentation, smoking, or other processes to enhance flavour or improve preservation. Most processed meats contain pork or beef, but might also contain other red meats, poultry, offal (e.g. liver), or meat byproducts such as blood. (IARC 2015, p. 1599)

This contradiction between the CRUK and IARC categorization, as well as the slightly tentative language used by the IARC (as evidenced, for example, by the weak modal *might* in the excerpt reported above), is only partially surprising: as we see in Section 4 of this paper, the semantic scope of *meat* is at least partially open to interpretation, because it is culture-dependent. This uncertainty about what counts as what type of meat is also visible from the

top-left-hand area of Figure 1, where *salami*, *bacon*, and *sausages and hot dogs* are shown in the *processed meats* grouping, but the use of emphatic caps for the word *includes* suggests that there might be more items in this group.

Lexical issues like the correspondence between countable nouns identifying animals and mass nouns standing for their meat are actually quite widely studied in ecolinguistic literature, especially in the branch that looks at the discursive representation of animals, not only as food, but also as pets, pests, and other, often adopting a corpus assisted discourse analysis approach (Cook 2015; Gilquin, Jacobs 2006; Pak, Sealey 2015). This area of study has focused quite closely over the years on a variety of texts about the natural world, including animals, showing a tendency, especially but not exclusively in scientific and agribusiness registers, to use mass nouns, often in the experiential role of Classifiers, to describe animals used for food (e.g. meat, poultry, venison, fish). This implicitly reinforces notions whereby animals, or at least those we eat, are "mere tonnage of stuff" (Stibbe 2014, p. 595; Fusari 2017, p. 140; Fusari 2018a, p. 297-304). Scholars pursuing this strand of ecolinguistics, endeavouring to investigate the human understanding – or, sometimes, *mis*understanding – of environmental issues, including meat eating, typically claim to follow in the footsteps of M.A.K. Halliday, specifically his keynote address delivered at the 1990 World Congress of the International Association of Applied Linguistics (AILA). Here, Halliday denounced the existence of "a syndrome of grammatical features which conspire […] to construe reality in a […] way that is no longer good for our health as a species" (Halliday 1990, p. 193). This discursive construction conveys, or even "engrammatizes" (Halliday 1990, p. 198), the idea that environmental resources, including the animals we eat, are inexhaustible, and can be tapped indiscriminately to accommodate progressive human demographic, economic and industrial growth. Halliday concluded that "the semantics of growthism" is a kind of hegemonic discourse, just like classism and sexism, and that it is a problem for biologists and physicists just as much as it is for linguists (Halliday 1990, p. 199). In the specific case of meat, the "ethics of semantics" adopted when talking about it is considered a problem also by animal industry professionals, especially those in charge of taking decisions about how to provide consumers with information about "the processing stage, in which cattle are transformed to beef and chickens become broilers or roasters, breasts, and even more vaguely, nuggets" (Croney, Reynnels 2008, p. 389). This has a clear impact on the ideologies embedded in scientific discourse, as "the 'black-boxing' that is entailed by nominalization might indicate an acceptance of the proposition as no longer requiring discussion" (Hunston 2013, p. 626).

Vague language is especially favoured by the discourse of agribusiness industry, to spare the consumer the most gory details of animal production, in what is an instance of discursive erasure (Stibbe 2012), a set of discourse strategies often used not only by the meat industry, but also by various other businesses that use animals (Fusari 2017).[5] Through discursive erasure patterns, animals are effectively removed from public consciousness, e.g. by not discussing them at all, by referring to them through euphemisms, or most typically by "treating the living world in the same discursive way as a stock of objects" (Stibbe 2015, p. 152). While this contributes to making animal industry practices more socially acceptable, and to "calming down the consumers" (Domingo, Nadal 2016, p. 114) in relation to the environmental and ethical impact of meat eating, including potential diet-based health issues (Packwood Freeman 2009), it also tends to drastically simplify reality. For example, unsettled terminological issues include which animals are good to eat (e.g. horse meat is considered a very nutritious type of red meat in some countries, while it elicits disgust in others), which provide red, white and dark meat, and what exactly is meant by *processing*, *carcinogen* or *carcinogenic* (Vicentini, Grego 2018, p. 362). Even the nature of tinned products, like Spam, as either meat or as an entirely different semantic category, is a matter of debate. These classifications are rooted in cultural differences, and can be explained in terms of historical motivations (Montanari 1993; Rodriguez-Wittmann 2014), but they also have a crucial impact on the scope and representativeness of epidemiological studies, which are obviously influenced by the way members of research teams class meat and animal types in their respective languages and cultures. This is why the contribution of linguistic studies can prove fundamental for a better understanding of meat carcinogenicity.

3. Methodology

For this study, a combined corpus assisted discourse analysis methodology was adopted (Partington *et al.* 2013), first on an unannotated corpus, and subsequently by tagging the corpus both for Parts of Speech using the TreeTagger engine developed by Schmid (1994), and for some Systemic Functional grammatical categories, especially those related to experiential

[5] Stibbe has stressed many times that discursive erasure is a pervasive phenomenon, extending well beyond nominalization and euphemism, and reaching into more complex issues of human consciousness as reflected in language. Although, for these reasons, a univocal definition of erasure is rather difficult to provide, it can be defined as "a story in people's minds that an area of life is unimportant or unworthy of consideration. An erasure pattern is a linguistic representation of an area of life as irrelevant, marginal or unimportant through its systematic absence, backgrounding or distortion in texts" (Stibbe 2015, p. 146).

meanings expressed in Transitivity patterns. This kind of grammatical tagging, based on a systemic formalism, can now be performed in a semi-automatic way, on a specialized corpus program developed by Mick O'Donnell at the Universidad Autonoma de Madrid, the UAM Corpus Tool (O'Donnell 2011). However, quite a lot of manual editing is still required, due to the presence of a physiological rate of error in the software output. This error rate should not be considered to be a limitation of the UAM software, but it is best viewed as intrinsically connected with the multi-tiered nature of Systemic Functional Linguistics, which entails the frequent conflation of functions in a single element, and is also sometimes open to interpretation, or multiple possibilities, in the grammatical labelling that can be associated with each phraseological pattern (Fusari 2016, p. 249). The deriving complications for corpus tagging can be more or less severe depending on the level of detail, or "delicacy" (Halliday, Matthiessen 2004, p. 45) that the researcher aims to achieve, but they cannot be fully eliminated, making the Systemic Functional grammatical formalism by far the most difficult to tag automatically in a corpus (O'Donnell 2005; O'Donnell, Bateman 2005; Fusari 2016). It is also, however, the most rewarding type of tagging for a meaning-centred analysis like the one performed in this study, because "the labelling of [Systemic Functional] grammatical features provides an interface to analysis at higher levels of abstraction that formal markup cannot, and does not aspire to, achieve" (Bartlett, O'Grady 2017, p. 6). These higher levels may include rhetorical strategies, which previous research has identified as being central to the way scientific beliefs about health issues are socially constructed, both within the relevant discourse communities, and among the general public (Arluke, Cleary, Patronek, Bradley 2018, p. 218). The lexicogrammatical features of scientific language (Halliday, Martin 1993) that contribute to construct the discursive reality of meat carcinogenicity are therefore better analyzed within a social semiotic (Halliday 1978, p. 2; Hestbaek Andersen, Boeriis, Maagerø, Seip Tonnessen 2015; Matthiessen 2017) than a structurally oriented methodological framework.

The corpus used for this study was assembled from *Elsevier Science Direct*, a large database of research journals available on subscription, looking for the keywords *IARC*, *meat,* and *cancer* in the sections:

- Agricultural/ Biological Sciences;
- Biochemistry;
- Genetics/ Molecular Biology;
- Environmental Science;
- Medicine/ Dentistry;
- Nursing/ Health Professions.

The search was made in August 2017, retrieving 39 articles, for a total of 384,491 words, complete with all references and appendixes. Files were then converted from PDF to TXT for perusal on Antconc (Anthony 2014).

The analysis related in this paper was also preceded by a pilot study (Fusari 2018a), performed in December 2015, which proved instrumental to the identification of the research questions worth addressing in subsequent studies such as the present one. The pilot study analyzed the discursive reactions to the IARC report not only from scientific sources, but also from animal rights ones, showing a high degree of intertextuality and register hybridity in the discursive construction of this scientific fact. The pilot study involved a much more restricted set of data (just below 50,000 words), to facilitate a manual close reading of all the texts, and as a way to test the reliability of automatic or semi-automatic corpus analysis, especially in relation to Systemic Functional tagging. The features analyzed (vocabulary, grammatical metaphor, and aspects of evaluative language, or Appraisal) showed that animal rights sources tended to appropriate the typical features of scientific language, mainly nominalization and the experiential structure of the noun group more generally (Bloor, Bloor 2013, pp. 140-148), to increase their credibility and claim for objectivity, while in fact their discursive aim was not so much informative as it was persuasive, i.e. trying to exploit the IARC communiqué to strengthen their arguments to convince people to go vegan. However, not even the scientific texts analyzed in our pilot study (two medical and one in the area of natural science) were devoid of ideological import, as they were not particularly concerned with the carcinogenicity of meat *per se*, or with the intrinsic truthfulness of the IARC findings, but largely evaluating the adherence of the IARC and other studies to the methodology of modern science, and therefore also their respect of the discursive order of science.[6] As these results seemed very promising not only for discourse analysis, but also for the study of genre integrity, hybridization, colonization, bending and mixing (Garzone 2015, p. 685), a decision was taken to expand the corpus, starting from scientific articles, as reported in this paper, and leaving animal rights texts as a potential development for further research.

4. Data and discussion

Given the focus of this study, and the keywords that were retrieved from *Elsevier Science Direct* to build the corpus, it is unsurprising that *cancer* and

[6] This is quite typical of the contemporary language of science, as "the method of science is realized again in the discourse it uses" (Tribble 2017).

meat are the two most frequent lexical words, and that the most recurrent premodifiers of *meat* are *red* and *processed*, in the noun group *red and processed meat*, referring intertextually to the title of the IARC report under discussion. What is more interesting is that there seems to be a denotational difference between the conventional use of *meat* as an uncountable mass noun, and its plural, *meats*, which in these texts identifies meat that is in some way unconventional. These *meats* may include less typically eaten animals (e.g. ostrich), and meat substitutes for vegetarians, vegans, as well as for those increasing numbers of people who are not strict vegetarians, but try to reduce the amount of meat they eat, and describe themselves (or, to be more accurate, are described by nutritionists, e.g. Graça, Oliveira, Calheiros 2015, p. 87) as 'flexitarians'.

The relation between the two most frequent lexical words, *cancer* and *meat*, is consistently (332 hits) framed as one of *causal/ clear/ neutral/ overall/ convincing/ positive/ potential/ reported/ (non/no/statistically) significant/ strong/ weak association*, evaluated not so much against hard data, but more often in a set of 'meta-analyses', i.e. secondary analyses of previous observational or statistical studies, including, but not limited to the IARC's. This emphasis on observing a cancer-meat *association* (more rarely, *connection*) in the scientific literature confirms findings from our pilot study, referenced in Section 3 of this paper, showing that the main concern of the articles is not meat carcinogenicity in itself: the focus is rather on the adherence of the IARC and other studies to the methodology of modern science, and therefore also on their respect of the discursive order of science more broadly defined. In this sense, the intrinsic truthfulness of the IARC findings, or the extent to which they should revolutionize the public's eating habits to protect them against cancer risk, is beside the point: what matters is the rigour of the scientific analysis provided, as well as the soundness of its methodological approach.

Table 1 illustrates the 20 most frequent lexical words in our corpus,[7] highlighting the articles' tendency towards nominalization (e.g. *consumption*, *intake*) and intertextual reference to the IARC and other research (e.g. *study/ies*, *analysis*) which may reveal a scientifically significant association (e.g. *associated*) between eating meat, especially in large amounts (e.g. *high*), and developing cancer, especially certain types (e.g. *gastric*).

[7] The query was restricted to nouns and adjectives to facilitate the identification of patterns of nominalization and agency in ideational experiential analysis. As shown by the presence of two forms of the word *study* in Table 1, the corpus is still unlemmatized at this stage in the project.

Rank	Freq	Word
1	3469	cancer
2	3207	meat
3	1690	food
4	1442	risk
5	1161	consumption
6	985	health
7	959	red
8	815	studies
9	755	dietary
10	755	study
11	722	processed
12	713	analysis
13	703	intake
14	677	products
15	663	human
16	628	high
17	618	diet
18	553	gastric
19	514	associated
20	499	total

Table 1
Most frequent nouns and adjectives.

Although frequency alone, as illustrated in Table 1, is not necessarily revealing of the nature of texts (Baron, Rayson, Archer 2009), calculating word frequency is a good starting point for most corpus analyses, especially for fairly small specialized corpora like the one under analysis here, as it tends to highlight both the topics that are most frequently mentioned in the texts, and the order of magnitude of the data at hand (Fusari 2018b, p. 6). To interpret frequency correctly, it is, however, always necessary to take a step further towards investigating patterns of use, as evidenced in collocations, word clusters and concordances.

The collocate list of *meat* (Table 2), obtained with Mutual Information[8] and sorted by frequency, also shows the repetitive behaviour of some keyword clusters (e.g. *red and/ or processed meat consumption*), as well as the most important terms associated with *meat* in the articles, which are indicative of the authors' interest in the semantic relation between meat, its colour (e.g. *red*, *white*), and the animals that may provide it. These include not only *poultry*, *fish* and *beef* (which appear in Table 2 as the most frequent), but also *pork*, *lamb* and *goat*. All these animals are classed as providing red meat, except in two articles (Domingo, Nadal 2017; Lippi *et al.* 2016), which

[8] Mutual information is one of the most commonly used statistical collocation extraction techniques. Specifically, "mutual information is the quantity that measures the mutual dependence of the two words/word combinations" (Metin, Karaoğlan 2011, p. 177).

mention animal age as a factor leading to the potential classification of young pigs as providing *white*, rather than *red* meat, possibly as a result of lower heme iron concentration in their muscles.

Rank	Freq	Freq L	Freq R	Stat (MI)	Collocate
1	1659	707	952	389.551	and
2	1290	1052	238	346.591	of
3	1050	967	83	703.637	red
5	758	379	379	482.463	meat
6	710	571	139	688.140	processed
7	613	170	443	598.417	consumption
8	359	50	309	599.041	products
9	326	141	185	448.613	or
10	225	154	71	575.647	total
11	210	54	156	516.245	intake
12	197	47	150	276.733	cancer
13	178	29	149	388.746	risk
14	100	3	97	522.751	science
15	96	88	8	698.275	artificial
16	79	29	50	622.068	poultry
17	75	24	51	412.878	associated
18	68	9	59	528.569	fish
19	68	28	40	436.734	beef
20	67	22	45	488.952	quality

Table 2
Collocates of *meat*, first 20 hits sorted by frequency.

Clusters on left and cluster on right (Table 3 and 4) of *meat* show that, while most articles[9] restrict their focus to the subject of the IARC report, i.e. red and processed meat, others open up the space for an extension of other animal-based foods that may play a role in cancer (e.g. *meat and meat products*, *meat and charcuterie*, *meat and fish*, *meat and dairy* in Table 3), or mention the existence of alternative meat-like products (some *artificial*, Table 4), which may be safer from the point of view of cancer risk, but could also, at the same time, satisfy the consumer's hunger for meat.

[9] The number of articles in which each cluster appears is shown in the column entitled 'Range'.

Rank	Freq	Range	Cluster
1	398	20	meat consumption
2	162	22	meat products
3	122	16	meat intake
4	104	12	meat and meat
5	81	8	meat science
6	41	4	meat quality
7	37	4	meat production
8	25	12	meat and processed
9	24	2	meat and charcuterie
10	21	8	meat and fish
11	19	2	meat quality traits
12	18	1	meat quintile
13	13	6	meat and dairy
14	11	5	meat cooking
15	11	5	meat processing
16	11	1	meat quartile
17	10	2	meat and pancreatic
18	10	2	meat substitutes
19	10	4	meat-based
20	10	5	meat and colorectal

Table 3
Word clusters with *meat* on left.

Rank	Freq	Range	Cluster
1	639	28	red meat
2	486	33	processed meat
3	86	1	artificial meat
4	85	20	of red meat
5	73	2	total red meat
6	49	2	cultured meat
7	45	5	total meat
8	43	2	white meat
9	34	6	consumption of meat
10	32	2	processed red meat
11	26	9	cooked meat
12	22	3	effects of meat
13	22	2	in vitro meat
14	21	5	cured meat
15	18	6	type of meat
16	17	3	eating meat
17	17	3	samples of meat
18	16	3	conventional meat
19	15	1	imitation meat
20	14	2	poultry meat

Table 4
Word clusters with *meat* on right.

The clusters also highlight a preoccupation with the amount and cooking methods of meat (i.e. *quintile*, *quartile* and *cooking* in Table 3; *total* and *cooked* in Table 4), as well as with the body organs for which the evidence of a relation between cancer and meat consumption is stronger (*pancreatic* and *colorectal* in Table 3).

Another issue that emerges from these data is whether *poultry* and *fish* count as meat: not all articles espouse the classification of poultry as a kind of meat, as some refer to *cooking or processing of white meat and poultry*, *enhanced/ high intake of white meat or poultry*, *fermented meat or poultry*, making a distinction between the two. Although the FAO/ WHO Food Standards Programme, in its Codex Alimentarius,[10] defines meat as "all parts of an animal that are intended for, or have been judged as safe and suitable for, human consumption" (Codex Alimentarius Commission 2005, p. 6), so potentially including not only birds, but also fish, consumer perceptions are actually much more variable. Even the *Dietary Guidelines for Americans* (U.S. Department of Health and Human Services, U.S. Department of Agriculture 2015), an official US government report published every 5 years, is not entirely clear on whether *poultry*, *meat* and *fish* fall within the same subgroup of *total protein*, or qualify as distinct foods (McNeill, Belk, Campbell, Gifford 2017, p. 37).

Other areas of semantic uncertainty, as briefly seen above, extend to the association between animals and meat colours. Some articles are very explicit in stating that the colour of meat, as well as its breakdown into food types, e.g. *processed meat*, may vary greatly across cultures. Semantic ambiguities in this context include the reference of *charcuterie* (Table 3) and *cured meat* (Table 4) as including or excluding products that are made from the meat of chickens, and the status of *dark* meat as identifying some specific parts of animals (i.e. the thighs and legs of chicken, turkey and fowl) which actually qualify as providing *white meat* when the body of the animal is considered in its entirety. This semantic debate has an obvious impact also on the evaluation of the degree of cancer risk that each of these meat types or products may pose, as what exactly counts as what meat colour remains open to interpretation.

Cooking methods are also brought into cause by some of the texts in this corpus, as a culture-specific variable that may affect not only the understanding, but also the degree of carcinogenicity of different types of meat and meat products. For example, a particular combination of deep frying, high cooking temperatures, food drying, and spice use by Indians is mentioned in one article (Gandhi *et al.* 2017) as being implicated in raising the rate of

[10] The Codex Alimentarius is a collection of food safety standards developed by a joint FAO/WHO Commission established in 1963. The full list of Codex standard and guidelines is available on the FAO website: http://www.fao.org/fao-who-codexalimentarius/codex-texts/all-standards/en/.

stomach cancer, despite low red meat intake. This suggests that consumer behaviour and cultural practices, like cooking, may be as important as food choices in terms of health. In fact, although the wide majority of the texts in this corpus are clinical studies, they exhibit a constant preoccupation not only with how consumers cook their food, but also with how they think and feel about it, as shown in the concordance in Table 5.

1	raises an important problem of acceptance by **consumers**. A third route for the future is simply
2	should not mean adverse health effects for the **consumers**, a number of issues (e.g., specific fish and
3	50% uptake seems unlikely to be acceptable to **consumers**. Consumer acceptability barriers in some
4	global adoption of insects as a food source is **consumer** acceptability (Looy et al., 2013; Shelomi,
5	and Schlüter, 2013). But issue of limited **consumer** acceptability is prevalent particularly in western
6	, Roosen, J., & Bieberstein, A. (2014). **Consumer** acceptance of new food technologies: Causes and
7	the objective messages to society. 5.3. **Consumer** food purchasing behavior. Consumer acceptance
8	appearance and aroma, and having high **consumer** acceptance (http:// www.likemeat.eu/). These
9	D. (2015). Impact of terminology on **consumer** acceptance of emerging technologies through the
10	. However, some experts showed that **consumer** acceptance of meat substitutes depends mainly on
11	factor in purchase decisions. Without **consumer** acceptance, otherwise appropriate food processing
12	sensory congruence issues and good **consumer** adhesion. In the case of food additives and
13	ever, policymakers, researchers and **consumers** alike are often overwhelmed by the complexity of
14	ally significant differences between **consumers** and non-consumers of these meats in case-control
15	matrices, which is important for both **consumers** and food manufacturers for producing healthier
16	the combination of food discarded by **consumers** and due to over-consumption halves from the
17	(Neu5Gc) into the tissues of red meat **consumers** and the subsequent interaction with inflammation
18	to note that, depending on the type of **consumer** and his/her expectations, it appears possible to
19	ally linked to the standard of living of **consumers** and is therefore of a financial nature which
20	eat product is hugely important to the **consumer**, and in some cases overrides fear of chemicals and

Table 5
Concordance of *consumer** (including plural), first 20 lines sorted right, first 20 lines.

Consumer *acceptance*/ *attitudes*/ *behavior(u)r*/ *choice*/ *demand*/ *expectations* etc. are actually seen as having an impact on all aspects of the relationship between meat eating and human health, from the possibility to market meat substitutes (both natural, like tofu, seitan and insects, and artificial, like in vitro/ cultured meat) to the scientific validity of epidemiological studies (for example, participants in research investigating eating habits are described as not being always accurate when they estimate the amount and type of food they eat), reaching to linguistic issues, both in doctor-patient communication (i.e. educating individuals to eat or avoid certain foods to live a healthier life) and in communication campaigns (i.e. information provided by government agencies to make scientific discoveries understandable by the general public, often through the filter of their specific policy priorities). The concordance in Table 5 can only provide a limited amount of context, for reasons of space, but it still manages to capture the multifarious dimensions of consumer acceptance in relation to meat eating, e.g. whether – often depending on their culture of origin – they will consider insects to be edible and to fall within 'meat' (line 4); how food technologies like GMOs or in vitro meat can work

for environmentally-minded consumers (line 6 and 9); and what role meat substitutes (line 10) and chemical additives (lines 12 and 20) may play in their choices. Overall, concordance data extracted from this corpus show that, in medical research papers, the consumer is not discursively constructed as a passive or impotent spectator of the meat/ cancer debate, but as a fundamental player in the response to new scientific narratives and discoveries about the relation between health and food more generally. This dialogic scenario, with consumers playing an active role in negotiating science through discourse, has also emerged in a recent linguistic study of the same IARC communiqué (Vicentini, Grego 2018), focusing on how meat carcinogenicity was reported to the general public both by the media and by scientific institutions.

As many as five articles in our own corpus mention *terminology* as having a direct impact on consumer acceptance of food and food processing technologies, showing quite a high level of metalinguistic awareness on the authors' part, including about the heterogeneity of meat definitions. Other articles dwell on the rhetorical strategies that could or should be used to either "calm down the consumers" (Domingo, Nadal 2016, p. 114), or to make them take action to reduce the amount of meat they eat, and it is suggested that this should be done by using "positive language" (Arena *et al.* 2017, p. 425), i.e. not by recommending that meat eating be avoided completely, like tobacco smoking,[11] but by presenting meat avoidance as an opportunity to try out new foods and enjoy a more varied diet. The use of metaphor is explicitly mentioned in one article as a potential communication tool to achieve this goal:

> While there are several metaphors to use to describe and explain actions once a person has been diagnosed with a chronic disease, there are very few metaphors to discuss the ways we prevent disease and promote HL[12] behaviors. Metaphors have profound influences on how people attempt to solve problems, particularly health problems. The ways in which we choose to message promotion of HL behaviors or prevention of chronic disease can have a profound effect on whether an individual is persuaded to act accordingly. The use of positive language or asset modeling are far superior in terms of prevention. (Arena *et al.* 2017, p. 425)

Another, even more basic metalinguistic issue that recurs in this corpus is the referent of meat itself, i.e., what we mean exactly by *meat*, and the denotational and connotational differences with other related words, like *muscle* and *protein*, as shown in the example below, taken from an article about the possibility to grow artificial meat from stem cells:

[11] The IARC has classed processed meat in the same group of carcinogens, 1A, as cigarettes, but this does not necessarily mean that meat and tobacco are equally dangerous, as explained in a FAQ list published by the World Health Organization in the wake of the publication of the IARC report under discussion (WHO 2015).

[12] The acronym HL stands for 'Healthy Living', as explained in the same article.

> The fact that artificial meat proponents have called their product "artificial meat" and not "artificial muscle" or even "artificial muscle proteins" (which would be more accurate) recognises implicitly that the word *meat* represents positive values: so, for example, meat is a symbol of force (inherited from the fact that primitive hunters had to be strong to hunt wild animals) and of high nutritional value (meat provides proteins in quantity and quality and many micro-nutrients which are beneficial for health) […] In fact, meat is a widely-consumed food in the world in different forms (fast cooking, slow cooking, ready-prepared meals, cured meats etc.), which shows how popular it is. In reality […] the product which is produced by stem cell culture is, from a strictly technical and semantic point of view, muscle tissue, (and even this point can be debated) and not meat. (Hocquette 2016, p. 169)

However, perhaps the clearest indication of how important it would be to achieve a set of universally agreed definitions of meat subtypes is in one of the very first articles that appeared in an *Elsevier Science Direct* journal in the wake of the IARC communiqué (Lippi *et al.* 2016). This article, an oncology paper published as a pre-print in 2015, and also investigated as part of our pilot study (Fusari 2018a) mentioned in Section 3 above, describes setting the "semantic debate" on what animals correspond to what meat colour, and providing a decisive definition of "processing", as "unavoidable steps in future clinical studies aimed to investigate the association between meat consumption and cancer" (Lippi *et al.* 2016, p. 12). The contribution linguists could make towards taking these steps is no doubt fundamental, and it is a challenge for further research, in both medicine and linguistics together, as we see in our conclusions.

5. Conclusion

This paper has investigated the discursive construction of meat carcinogenicity through a case study of a series of scientific articles that were published shortly after the release of the announcement that the IARC had placed red and processed meat in its list of cancer-causing agents. The IARC's is not the only existing classification standard for carcinogens,[13] but it is very well reputed throughout the scientific community, so this statement had a tremendous impact both on science and in the media, arousing, at the same time, interest and controversy (Kelland 2016).

Although the sources we have examined show that scientists are very preoccupied with consumer acceptance when making recommendations and drawing conclusions about eating habits, it appears quite clear that the aim of these publications is *not* to inform consumers about the potential health hazard of eating certain types of meat: the addressees of the articles under

[13] Other standards have been developed mainly in the United States, by the Occupational Safety and Health Administration (OSHA), the American Conference of Governmental Industrial Hygienists (ACGIH) and the National Toxicology Program (NTP).

investigation are other scientists involved in cancer studies, who are co-constructing the connection between various types of meat-based foods and the development of cancer in humans. In doing so, scientists are providing each other with references and evaluations, often in the form of meta-analyses, which are necessary to make a scientific claim acceptable by the methodological and discursive conventions of the scientific community. This is why, in communicating with the general public, the WHO has been much more explicit than the IARC has been in its original report (IARC 2015), stating that "the latest IARC review does not ask people to stop eating processed meats" (Härtl 2015). Such an invitation is never extended in the articles in our corpus, as it would simply fall outside the scope of scientific literature.

Scientific literature, as exemplified in the small corpus we have investigated, does not actually aspire to provide some "silver bullet" truth[14] that will settle a given matter definitively, either for the public and or for other scientists. Its aim is rather to "persuade readers [i.e. other members of the relevant discourse communities] of the scientific acceptability of the knowledge claims presented" (Allen *et al*. 1994, p. 280), especially through the rhetorical instruments of cross-reference and evaluation. It would certainly be unfair, and perhaps also grossly misplaced, to state that the only, or even the main preoccupation of scientific literature is rhetorical: however, at the same time, it is undeniable that rhetoric plays a vital role in linguistically constructing the reality of modern science, both in terms of metadiscourse (Hyland 2017) and in more fundamentally grammatical ways (Halliday 1989).

As concerns the studies assembled in our corpus, to make it even clearer that their fundamental concern is not the intrinsic truthfulness of their findings, but their adherence to the methodology and discursive order of modern science, the WHO (2015) has explicitly addressed one specific issue raised by the general public, i.e. the fact that processed meat has been placed in the same category of carcinogens as cigarettes and asbestos:

> Processed meat has been classified in the same category as causes of cancer such as tobacco smoking and asbestos (IARC Group 1, carcinogenic to humans), but this does NOT mean that they are all equally dangerous. The IARC classifications describe the strength of the scientific evidence about an agent being a cause of cancer, rather than assessing the level of risk. (WHO 2015, p. 9)

The difference between "strength of scientific evidence" and "level of risk" in the excerpt above may be less than clear for a non-specialist audience (as the public may rightfully believe that there is a cause-effect relation between

[14] On the problematic and multiple notions of truth in medical writing, see Skelton 1997.

the two) but it is a very important distinction in terms of the discursive construction of a scientific fact.

Our study has highlighted another area of semantic uncertainty that raises fundamental questions about how consumers should be informed about the level of health risk involved in meat eating, i.e. the exact meaning of the word *meat*. Some examples shown in Section 4 of this paper have actually revealed quite an amazing level of metalinguistic awareness on the part of medical scholars writing about the IARC report, specifically in conjunction with a set of culture specific issues, e.g. the colours, ethical values, and cognitive metaphors that are associated with eating animals. The branches of linguistics that are explicitly mentioned in this corpus as capable of making a useful contribution to medical research on the meat-cancer relation run the gamut of our fields of study, including terminology, metaphor, rhetoric, and ethnolinguistics.

Perhaps the most stimulating development for further research in this area would consist in taking up the challenge launched by one of the articles collected in this corpus (Lippi *et al.* 2016), which quite openly calls upon linguists to help epidemiologists and clinicians develop a set of universally accepted definitions of meat and of its various byproducts. Such definitions are expected to be instrumental in overcoming the multiple issues of culture-specificity that make the existing terminology databases and taxonomies in this branch of medical studies still largely inconclusive and less than comprehensive.

However, the very fact that these concepts are specific to different cultures may actually make it rather complicated to reach a universal terminological agreement, and especially to bring it home to the general public, who are likely to continue thinking of meat in terms of the associations it has in their cultures.

Bionote: Sabrina Fusari holds a PhD in Intercultural Communication and is Associate Professor of English Language and Linguistics at the Department of Modern Languages, Literatures and Cultures of the University of Bologna (Italy). She teaches Systemic Functional Linguistics and Corpus Linguistics to both undergraduate and graduate students, with a focus on ideational experiential meaning as expressed in a variety of registers of English, both as a native language, and as a lingua franca. Her research interests include critical discourse analysis, intercultural rhetoric, ecolinguistics, media discourse, and English for specific purposes.

Author's address: sabrina.fusari2@unibo.it

References

Allen B., Qin J. and Lancaster F.W. 1994, *Persuasive communities: a longitudinal analysis of references in the Philosophical Transactions of the Royal Society, 1665-1990*, in "Social Studies of Science" 24 [2], pp. 279-310.

Anthony L. 2014, *AntConc (Version 3.4.4)*. http://www.laurenceanthony.net/ (19.01.2018).

Arena R., McNeil A., Sagner M. and Hills A.P. 2017, *The current global state of key lifestyle characteristics: health and economic implications*, in "Progress in Cardiovascular Diseases" 59, pp. 422-429.

Arluke A., Cleary D., Patronek G. and Bradley J. 2018, *Defaming Rover: error-based latent rhetoric in the medical literature on dog bites*, in "Journal of Applied Animal Welfare Science" 21 [3], pp. 211-223.

Austin J.L. 1962, *How to Do Things with Words*, Oxford University Press, Oxford.

Baron A., Rayson P. and Archer D. 2009, *Word frequency and keyword statistics in historical corpus linguistics*, in "Anglistik: International Journal of English Studies" 20 [1], pp. 41-67.

Bartlett T. and O'Grady G. (eds.) 2017, *The Routledge Handbook of Systemic Functional Linguistics*, Routledge, London.

Bloor T. and Bloor M. 2013, *The Functional Analysis of English*, Routledge, London.

Codex Alimentarius Commission 2005, *Code of Hygienic Practice for Meat*. http://www.fao.org/fao-who-codexalimentarius/sh-proxy/en/?lnk=1&url=https%253A%252F%252Fworkspace.fao.org%252Fsites%252Fcodex%252FStandards%252FCAC%2BRCP%2B58-2005%252FCXP_058e.pdf (21.01.2018).

Cook G. 2015, *"A pig is a person" or "You can love a fox and hunt it:" innovation and tradition in the discursive representation of animals*, in "Discourse and Society" 26 [5], pp. 587-607.

Croney C. and Reynnells R. 2008, *The ethics of semantics: do we clarify or obfuscate reality to influence perceptions of farm animal production?*, in "Poultry Science" 87 [2], pp. 387-391.

Domingo J. and Nadal M. 2016, *Carcinogenicity of consumption of red and processed meat: what about environmental contaminants?*, in "Environmental Research" 145, pp. 109-115.

Domingo J. and Nadal M. 2017. *Carcinogenicity of consumption of red meat and processed meat: A review of scientific news since the IARC decision*, in "Food and Chemical Toxicology" 105, pp. 256-261.

Dunlop C. 2015, *Processed meat and cancer – what you need to know*. http://scienceblog.cancerresearchuk.org/2015/10/26/processed-meat-and-cancer-what-you-need-to-know (23.01.2018).

Fusari S. 2016, *The role of corpus annotation in the SFL-CL marriage: A test case on the EU debt crisis*, in Gardner S. and Alsop S. (eds.), *Systemic Functional Linguistics in the Digital Age*, Equinox, Sheffield, pp. 246-259.

Fusari S. 2017, *What is an animal sanctuary? Evidence from applied linguistics*, in "Animal Studies Journal" 6, pp. 137-160.

Fusari S. 2018a, *"Bacon wrapped cancer": The discursive construction of meat carcinogenicity*, in "Text & Talk" 38 [3], pp. 291-316.

Fusari S. 2018b, *Changing representations of animals in Canadian English (1920s-*

2010s), in "Language & Ecology" 3-4, pp. 1-32.
Gandhi A.K., Kumar P., Bhandari M., Devnani B. and Rath G.K. 2017, *Burden of preventable cancers in India: Time to strike the cancer epidemic*, in "Journal of the Egyptian National Cancer Institute" 29, pp. 11-18.
Garzone G. 2015, *Genre analysis*, in Tracy K. (ed.), *The International Encyclopedia of Language and Social Interaction*, Wiley Blackwell, Oxford, pp. 677-693.
Gilquin G. and Jacobs G. 2006, *Elephants who marry mice are very unusual: the use of the relative pronoun who with nonhuman animals*, in "Society and Animals" 14 [1], pp. 79-105.
Graça J., Oliveira A. and Calheiros M.M. 2015, *Meat, beyond the plate: data-driven hypotheses for understanding consumer willingness to adopt a more plant-based diet*, in "Appetite" 90, pp. 80-90.
Halliday M.A.K. 1978, *Language as Social Semiotic. The Social Interpretation of Language and Meaning*, University Park Press, Baltimore.
Halliday M.A.K. 1989. *Some grammatical problems in scientific English*, in Halliday M.A.K. and Martin J.R. (eds.) 1993, *Writing Science: Literacy and Discursive Power*, University of Pittsburgh, Pittsburgh, pp. 69-85.
Halliday M.A.K. 1990, *New ways of meaning: the challenge to applied linguistics*, in Fill A. and Mühlhäusler P. (eds.), 2001, *The Ecolinguistics Reader*, Continuum, London, pp. 175-202.
Halliday M.A.K. and Martin J.R. (eds.) 1993, *Writing Science: Literacy and Discursive Power*, University of Pittsburgh, Pittsburgh.
Halliday M.A.K. and Matthiessen C.M.I.M. 2004, *An Introduction to Functional Grammar. Third Edition*, Arnold, London.
Härtl G. 2015, *Links between processed meat and colorectal cancer. WHO statement*. http://www.who.int/mediacentre/news/statements/2015/processed-meat-cancer/en/ (24.01.2018).
Hestbaek Andersen T., Boeriis M., Maagerø E. and Seip Tonnessen, E. 2015, *Social semiotics. Key figures, new directions*, Routledge, London.
Hocquette J.F. 2016, *Is in vitro meat the solution for the future"?*, in "Meat Science" 120, pp. 167-176.
Hunston S. 2013, *Systemic functional linguistics, corpus linguistics and the ideology of science*, in "Text & Talk" 33 [4-5], pp. 617-640.
Hyland K. 2017, *Metadiscourse: what is it and where is it going?*, in "Journal of Pragmatics" 113, pp. 16-29.
IARC 2015, *Carcinogenicity of consumption of red and processed meat*, in "The Lancet Oncology" 16 [16], pp. 1599-1600.
Kelland K. 2016, *Who says bacon is bad? How the World Health Organization's cancer agency confuses consumers*. https://www.reuters.com/investigates/special-report/health-who-iarc/ (24.01.2018).
Lippi G., Mattiuzzi C. and Cervellin G. 2016, *Meat consumption and cancer risk: a critical review of published meta-analyses*, in "Critical Reviews in Oncology/Hematology" 97, pp. 1-14.
Matthiessen C.M.I.M. 2017, *Language use in a social semiotic perspective*, in Barron A., Yueguo G. and Steen G. (eds.), *The Routledge Handbook of Pragmatics*, Routledge, London, pp. 459-489.
McNeill S.H., Belk K.E., Campbell W.W. and Gifford C.L. 2017, *Coming to terms: meat's role in a healthful diet*, in "Animal Frontiers. The Review Magazine of Animal Agriculture" 7 [4], pp. 34-42.

Metin S. K. and Karaoğlan B. 2011, *Measuring collocation tendency of words*, in "Journal of Quantitative Linguistics" 18 [2], pp. 174-187.

Montanari M. 1993, *La fame e l'abbondanza: storia dell'alimentazione in Europa*, Laterza, Bari.

Montanari M. 2015, *Medieval Tastes: Food, Cooking and the Table*, Columbia University Press, New York.

National Geographic Education 2015. *What the World Eats*. https://www.nationalgeographic.com/what-the-world-eats/ (19.01.2018).

O'Donnell M. 2005, *The UAM systemic parser*, in "Proceedings of the 1st Computational Systemic Functional Grammar Conference, 16 July 2005, Sydney, Australia". http://www.wagsoft.com/Papers/ODonnellUamParser.pdf (19.01.2018).

O'Donnell M. 2011, *UAM Corpus Tool (Version 2.8.).* http://www.corpustool.com (19.01.2018).

O'Donnell M. and Bateman J. 2005, *SFL in computational contexts: a contemporary history*, in Webster J.J., Hasan R. and Matthiessen C.M.I.M (eds.), *Continuing Discourse on Language: A Functional Perspective*, Equinox, London, pp. 343-382.

Packwood Freeman C. 2009, *This little piggy went to press: the American news media's construction of animals in agriculture*, in "The Communication Review" 12 [1], pp. 78-103.

Pak C. and Sealey A. 2015, *"An urban fox is a bushy-tailed James Dean, living fast and dying young:" representations of foxes in UK discourse*. https://animaldiscourse.wordpress.com/?page_id=459 (19.01.2018).

Partington A., Duguid A. and Taylor C. 2013, *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*, John Benjamins, Amsterdam.

Rodriguez-Wittmann K. 2014, *Mala digestio, nulla felicitas: gastronomy as an element of well-being in the Tacuinum Sanitatis*, in Barrera N., Pellissa-Prades G., Nieto-Isabel D., Sallés Vilaseca L., Rabassó G., Elies I. and Bellver J. (eds.), *Spaces of Knowledge. Four Dimensions of Medieval Thought*, Cambridge Scholars, Newcastle-upon-Tyne, pp. 25-34.

Schmid H. 1994, *Probabilistic part-of-speech tagging using decision trees*, in "Proceedings of the International Conference on New Methods in Language Processing, 1994, Manchester, UK". ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf (19.01.2018).

Skelton J. 1997, *The representation of truth in academic medical writing*, in "Applied Linguistics" 18 [2], pp. 121-140.

Stibbe A. 2012, *Animals Erased: Discourse, Ecology, and Reconnection with the Natural World*, Wesleyan University Press, Middletown.

Stibbe A. 2014, *Ecolinguistics and erasure: Restoring the natural world to consciousness*, in Hart C. and Cap P. (eds.), *Contemporary Critical Discourse Studies*, Bloomsbury, London, pp. 583-602.

Stibbe A. 2015, *Ecolinguistics: Language, Ecology and the Stories We Live By*, Routledge, Abingdon and New York.

Tribble C. 2017, *Expert or native? Unpacking conflicting paradigms in EAP writing instruction*, Plenary lecture delivered at the *2017 CLAVIER International Conference, Representing and Redefining Specialised Knowledge*, Bari, December 1, 2017.

U.S. Department of Health and Human Services and U.S. Department of Agriculture. 2015, *2015-2020 Dietary Guidelines for Americans. Eighth Edition*. https://health.gov/dietaryguidelines/2015/resources/2015-

2020_Dietary_Guidelines.pdf (19.01.2018).
Vicentini A. and Grego K. 2018, *"Meat gives you cancer." The popularisation of scientific news with public health relevance*, in "Lingue e Linguaggi" 26, pp. 357-376.
World Health Organization 2015, *Q&A on the Carcinogenicity of the Consumption of Red meat and Processed Meat*. http://www.who.int/features/qa/cancer-red-meat/en/ (20.01.2018).

MAPPING MEDICAL ACRONYMS[1]

ANNA LOIACONO, FRANCESCA TURSI

Abstract – Searches in the multimedia *House Corpus* reveal that, as well as a noun, the acronym MRI functions in the *House MD* series as an adjective and, albeit rarely, as a transitive verb and that, besides referring to equipment used in the MRI procedure and to the procedure itself, it is also used as a countable noun, often in the plural (MRIs), to refer to the scans so produced. By contrast, the entry in the online OED (Third edition) refers to MRI only as a noun and restricts its definition to a medical procedure and associated equipment. Given these characteristics, the *House Corpus* project has been an opportunity to investigate medical acronyms more completely and, in particular, to meet the challenge they represent for medical trainees when listening to spoken medical discourse. With the assistance of student annotators, every medical acronym in the *House Corpus* has now been indexed in terms of grammatical (countable/uncountable noun; adjective and verb) and functional categories (specific diseases; therapeutic/diagnostic procedures; equipment; test results; medical facilities, names of substances; anatomical parts and body states). Special care has been taken in the tagging process to include derivative and related forms (e.g. fMRI as well as MRI). As a result, the *House Corpus* now has a specific *Acronym Search* resource, a first step towards *Acronym Maps* that aggregate the various grammatical and functional categories into which a specific acronym falls. While a clear boon for medical English classrooms, such *Maps* support hunches about the nature and incidence of acronyms in spoken and written forms of medical discourse in English and, when compared to other languages such as Italian, highlight differences in abbreviatory strategies. The article concludes that greater consideration of specialised medical genres and contexts, especially those relating to spoken discourse, (Loiacono 2015, 2016, 2018) needs to be made in corpus studies than has been the case in the past.

Keywords: acronyms; acronym maps; abbreviatory strategies; spoken medical discourse.

1. Introduction

For students in their first years of medical studies in Italian universities coming to terms with the acronyms used in clinical care constitutes a problem. In fact, it would be more accurate to say that the problems acronyms constitute fall into a number of very different categories. The first of these relates to how best to learn them. Like it or not, learning acronyms is an essential part of the fluency in reading medical discourse in English that medical undergraduates are expected

[1] Sections 1, 4 and 5 were written by Anna Loiacono, Sections 2 and 3 by Francesca Tursi.

Doi Code: 10.1285/i9788883051531p93

to achieve in their first years of university study. However, medical students, too, have their own expectations about the learning of acronyms, one of which is that their teachers, and not just teachers of medical English, should guide them as regards *which* acronyms should be learnt. In addition, medical trainees expect to receive advice on *how* to go about learning them. Alas, where available, such guidance is often unsatisfactory. Students' questions about whether it is best to learn acronyms by reading medical texts, by consulting online glossaries or by simply listening to classroom lectures and noting what acronyms are used are likely to go unanswered. This is because the processes that relate to the acquisition and use of medical acronyms are far more complex than would appear to be the case at first sight. They raise many learning issues that require considerable research.

SEASON: 1 - Episode: 17 - Role Model - Scene: 07

CAMERON: "Eastbrook Pharmaceuticals are pleased to announce that Dr. Gregory House will present the latest research on their exciting new **ACE** inhibitor."
CHASE: You're making that up. That's Vogler's company.
CAMERON: Press release. Doing an address at the North American Cardiology Conference. [Chase looks at the screen from behind Cameron.]
CHASE: House never gives speeches. [House enters.]
HOUSE: But when I really believe in something... Gosh dang it, I've got a chance to make a difference here.
CHASE: You made a deal with Vogler?
HOUSE: It's all the rage. Everybody's doing it. [Chase gives House a petty, pouty look and goes to sulk in a chair. Cameron walks over to House.]
CAMERON: So, what's the deal? You get to keep all of us if you plug his products?
HOUSE: One speech, no biggie. Foreman's doing a bone marrow biopsy to check for cancer.
CHASE: Cancer? The Senator's got **AIDS**.
HOUSE: Cancer sounds better on a press release. I need you guys to rush the **ELISA** test for **HIV**.

Figure 1
Acronyms with different forms and functions in a clinical context.

Ways of tackling the various issues are described below in relation to the ongoing development of a specialised acronym resource. Combined with the tools already available in the *House Corpus* interface (Taibi *et al. this volume*), this allows specific acronym searches to be made in the *House MD* TV series thereby providing a partial solution to some of the problems students face. The *Acronym Search* resource identifies scenes, such as the one shown in Figure 1, in which the searched-for acronyms are highlighted in red making them easy to distinguish. The current stage of development responds, in part, to some of the requests for assistance that students make, in particular thanks to the inclusion of a scene-by-scene link-up between the transcribed text and the corresponding video episode that provides students with an efficient way of hearing how these acronyms are pronounced. Thus, besides helping to distinguish between initialisms like HIV, pronounced letter by letter, and true acronyms modelled on pre-existing words like ACE and AIDS or names like ELISA, students now have a resource that allows them to acquire confidence in their ability to identify

acronyms in fast discourse – whether, for example, Dr. House is talking about EMGs (electromyograms) or ENGs (electroneurograms). The resource thus relieves the pressures on teachers mentioned above by providing a support for acronyms to be learnt in an online *self-learning* context. In theory, this encourages students to use contextual clues to figure out the basic function of an acronym even when they are unsure of its precise meaning – a matter which, despite the reassuring results described below, requires further assessment and more research.

Recognising acronyms *as* acronyms in both written and oral discourse is, indeed, less than half the battle for medical students. A second order of problems relates to acronyms' use and functions in medical discourse. This includes awareness of the constraints on using acronyms in oral and written discourse. The question – *What does LP actually stand for Lumbar Puncture or Lipoprotein?* – highlights the well-known problem of acronyms' ambiguity in medical contexts and the need to be able to identify and interpret their meaning readily (Pakhomov 2002). This much-debated feature in the medical and information technology literature (Berlin 2013; Kuhn 2007; Stevenson *et al.* 2009) includes the potential for errors to arise when, for example, doctors use an ambiguous acronym in medical notes without further specification or contextualisation (Parakh *et al.* 2011). This has led to claims that resolving acronyms' ambiguity is of paramount importance. However, while the perils of acronyms may be relevant in later years of study (for example, when learning to write research articles), the ambiguity issue appears to be overstated at least as far as initial medical studies are concerned. The analysis carried out in the construction of the *Acronym Search* resource (see *Section 3* below) revealed that very few of the acronyms used in the *House MD* series are, in fact, ambiguous and that context helps to clarify their meaning. Hence, rather than on constraints, attention in the early years of medical study should perhaps focus more on the affordances that acronyms provide in medical communication.

When asked to write a summary of a *House MD* episode in English and to practise their skills of abbreviation in English (see *Section 4*), students come to realise that there are crucial differences in the way 'English' medical acronyms are used in their mother tongue (mostly Italian for our students), and English discourse. When used in Italian medical discourse, English acronyms, such as CT or MRI, are grammatically invariable, whereas this is not the case in English. Figure 2 highlights the utterance "*ER CT'd him*" retrieved from the *House Corpus* using the *Acronym Search* resource, a striking example of abbreviation possible with acronyms in English but whose brevity and simplicity cannot be matched in Italian. Contrary to the frequent claims that full forms are preferable to acronyms (Baue 2002; Brubaker, Brubaker 1999; Kuhn 2007; Pakhomov 2002; Parakh *et al.* 2011; Patel, Rashid 2009, Pottegård *et al.* 2014; Summers, Kaminski 2004; Walling 2001), such examples suggest

that the acronyms used in English medical discourse are often *more*, rather than *less*, acceptable than the forms from which they are derived. The term *CT scan* appears in the *House Corpus* in 25 different scenes, but *Computed Tomography scan*, its multi-word source, never appears. Moreover, contrary to what is often assumed regarding acronyms' derivation from multi-word sources, there is no corresponding full form for the verb form *CT'd*. Had it existed, it would presumably have been **Computed Tomographied*, a rather awkward term to handle in both written and spoken discourse.

SEASON: 6 - Episode: 13 - Moving the Chains - Scene: 04

THIRTEEN: 22-year-old male – 6'7", 310 pounds. Clearly has brain involvement. [looking at the video of Daryl hitting himself] The guy has no recollection of this entire incident.
HOUSE: Football player. Those are the ones that get hit in the head a lot, right?
CHASE: **ER CT'd** him. No concussion, no stroke, no cortical degeneration.
TAUB: And he had a full psych evaluation. He's not crazy.
HOUSE: So it's roid rage. You don't think they grow them that big naturally.
FOREMAN: **ER** also tested for steroids. He's clean.
HOUSE: Only proving that our guy got his hands on the good stuff.
FOREMAN: The negative test at least means steroids is less likely. We should discuss other possibilities.

Figure 2
Acronyms support processes of metonymy and lexicalisation.

The frequency with which acronyms undergo metonymic processes is a further issue when attempting to master the abbreviation practices that underpin medical discourse. *ER* appears in many episodes in the *House MD* series (a total of 83 scenes). However, it is only through specialised corpora and thanks to corpus-specific annotations (see *Sections 2* and *4*) that medical trainees can ask and find answers for an all-important question – in what ways do the *uses* made of English acronyms in *Italian* medical discourse differ from those of the very same acronyms when used in medical discourse in *English*? For example, *ER* and *MRI* may, in medical discourse in English, be references to specific hospital facilities and their location in a hospital. They may also be references to these facilities' functions, which includes the services they deliver and, as Figure 2 shows, the staff who work there. Italian cannot abbreviate in *this* way. In Italian, it is necessary instead to spell out these different functions, possibly with a reference to *il servizio MRI* for the facility and to *gli addetti all'MRI* or *i tecnici dell'MRI* for the personnel. *Section 2* illustrates how specialised corpora can provide a useful way of addressing these issues, while *Section 4* describes how medical trainees can support efforts to master 'metonymic abbreviation' – essential for efficient medical communication in English.

A third type of problem relates to acronyms' use in digital texts. This has to do, in particular, with the skills required when attempting to retrieve data from digital databases and the degree to which abbreviated forms (acronyms in particular) can be used to this end. Like their counterparts in universities in other parts of the world, Italian medical students are given free access to digital

resources but many students are reluctant to use them. In the case of medical students, this is hardly surprising. In the early years of study, formulating questions in a clinical context is a major part of clinical training (see Loiacono 2018, pp. 691-695, for PICO questions in digital healthcare). The question – *Did a digital search miss out vital data?* – highlights the need to understand and successfully judge the probability that information has been missed owing to the way in which database queries are formulated. Formulating such queries in a way that is consistent with the medical tradition of question-formulation is a relatively new issue in medical training but is emerging as major requirement in Italy and elsewhere (Schultz 2006).

SEASON: 3 - Episode: 19 - Act Your Age - Scene: 08

FOREMAN: The stroke was caused by a clot in her middle cerebral artery. Started her on **TPA**. It should dissolve the clot and hopefully prevent brain damage, but we won't know for sure until she regains consciousness.
HOUSE: Or she has another stroke. Arthritis, heart disease, why can't this kid act her age?
FOREMAN: **JRA** doesn't affect the blood, means the clot's a symptom of something else. [Cameron walks in.]
CAMERON: It's a symptom of polycythemia, she's fully hydrated and her blood's still thicker than pancake batter.
HOUSE: Well thick blood explains the stroke, could also have caused an autoimmune response, which would explain the **JRA** kicking into gear. But what explains the thick blood?

Figure 3
Embracing variation from expected conventions in digital searches.

Once again, specialised corpora are – potentially – a way of sensitising students in their early years of study to the relevance of this tricky digital issue. As Figure 3 shows, through highlighting (and comparisons with other scenes), it is possible to encourage students to reflect on the diversity and variation in the process of abbreviating with its many subtleties (see also *Section 4*). The typical capitalisation of acronyms may help distinguish Dr. House's *ACE inhibitors* from his encounters with *ace attorneys* and his use of *PAS* to indicate *p-aminosalicylic acid* (*Scene 32, Episode 13, Season 8*) from those where he pretends not to be able to speak English (*je ne parle pas anglais*, *Scene 10, Episode 21, Season 7*) but, as Figure 3 illustrates, in the case of a transcriber's slip-up, breaches of the capitalisation rule come to be highlighted. Such examples help students to become aware of, and to anticipate, part-lower case, part-upper case acronyms, as well as further variants of such 'standard' hybrids – not just *tPA (tissue plasminogen activator)* but also partial acronym forms, such as *t-plasminogen activator*. Training students to *predict* typical patterns of word abbreviation is essential if they are to feel confident about their use of digital resources. The quality of the search queries they undertake will ultimately depend on their understanding of how rules about 'standard' conventions come to be broken.

All this points to the need for medical students to contemplate written, oral and digital discourse in their studies of acronyms as well as the lexicogrammatical, discourse and digital aspects of the process of abbreviation

in medical discourse in English. This article does not attempt to explore these issues individually. Rather our concern is with developing a *single* research, teaching and learning framework that potentially allows all aspects of acronyms to be addressed and which can be extended at a later stage of research to cover all aspects of abbreviation in clinical care. This will allow a better focus on abbreviation as a process to be learnt, taught and thoroughly practised within English for Medicine courses (*Section 4*).

By 'framework', we mean an online resource that can be used in specific teaching and learning contexts to underpin references to, and illustrations of, descriptive models of abbreviation in medical and scientific discourse. Indeed, the ultimate goal of the research is not to produce an interface that detects every acronym in a specialised corpus. Rather, it is to build a corpus resource that allows the issue of mastering the *functions* of acronyms in clinical discourse to be approached in a way that meets the demands in Italian universities of medical training in English. As explained below, the current project is a first step in this direction. Indeed, in order to function fully it will eventually need to take genre, and the relationships between acronyms and specific medical genres, as well as other issues into consideration, all of which is further discussed in *Section 4.*

2. Materials and Methods

The research so far undertaken is reported in summary form in this Section. It relates to the very first stages of annotation of acronyms in the *House Corpus.* With its customisable interface and scene-based indexing of scripted oral medical discourse of an entire TV series, the *House Corpus* (Taibi *et al.*, this volume) provides a suitable basis for the development of an online *Acronym Search* resource that identifies acronyms and illustrates their role in clinical discourse. In the first now completed stage of the research, manual annotation of all medical acronyms in the *House MD* TV series was undertaken. Given the project's initial focus on medical acronyms, the students who carried out the annotation (see *Acknowledgements*) were asked to exclude (a) abbreviations, except for part-acronym, part-abbreviation compounds (such as A-fib = atrial fibrillation) and (b) acronyms with no clear medical reference (e.g. LA = Los Angeles). For each transcript, a *Summary Table* was produced that established the type/token ratio for each episode. In addition, each transcript was annotated with the functional and grammatical tags reproduced in Tables 1 and 2.

TAG	DEFINITION OF FUNCTIONAL TAGS
1. BODY PARTS	A part of the human body e.g. CNS - Central Nervous System
2. BODY STATES	Refers *not* to a DISEASE but to the current *state* of *part* of a patient's body that is not functioning correctly, which has suffered a *lesion*.
3. DISEASES	Refers to the name of a specific disease.
4. FACILITY	The *place* where a procedure (diagnosis/therapy) is carried out or equipment used.
5. METATEXTUAL	Acronyms explained: e.g. *House: DNR means "do not resuscitate", not "do not treat".*
6. PROCEDURE	Unlike a specific DIAGNOSTIC TEST, this is used to describe an *action* to be undertaken, or one already completed; this label is usually associated with a U-NOUN as it is more abstract
7. SUBSTANCE	Typically a drug introduced as part of a therapy/test: e.g. IgG in immunoglobulin therapy
8. TESTS	A diagnostic test still to be done or test result for an already completed test

Table 1
Functional tags for House Corpus acronyms.

These Tables were part of a short manual that the annotators were given to guide their annotation. The annotations made by the students effectively tested out the validity of the acronym model submitted to them.

TAG	DEFINITION OF GRAMMATICAL TAGS
1. U-NOUN	UNCOUNTABLE NOUN as in *MRI stands for Magnetic Resonance Imagining*; *MRI works wonders*. You can't physically touch these MRIs or count them up … so no singular and plural difference exists; they are therefore uncountable. Another example is: *It's ALS.*
2. SC-NOUN	SINGULAR COUNTABLE NOUN e.g. *an MRI: he's a DNR* (…. *She's another DNR*…). NB. a U-NOUN often "becomes" an SC-NOUN when preceded by an article, number, possessive or demonstrative adjective: in e.g. *my ALS?*
3. PC-NOUN	PLURAL COUNTABLE NOUN e.g. *two MRIs*; they typically have a lower case *s*
4. ADJ	ADJECTIVE which precedes the noun it qualifies *e.g. a DNR patient*.
5. PRED-ADJ	PREDICATE ADJECTIVE: used after a verb e.g. he's DNR (NB. no *a/the* etc).
6. VERB	ANY VERB FORM: *He needs MRI-ing*; *she's been MRI-ed*; *I want to MRI him.*

Table 2
Grammatical tags for House Corpus acronyms.

The student annotators were given the opportunity to indicate their doubts. In particular, they were instructed to use the annotational label UNDECIDED to indicate those cases where an acronym appeared not to comply with the definitions given for the grammatical and functional model supplied. In fact, very few such cases were reported. When analysed, they highlighted uncertainties between categories – whether, for example, an acronym related to a PROCEDURE or a TEST. Most of these cases were subsequently resolved,

often by the student annotators themselves, by comparing other instances of the same or similar acronyms in the various episodes.

Other doubts related to the absence of certain acronym categories from the model, for example, relating to healthcare personnel (e.g. *EMT = Emergency Medical Technician*) and healthcare administration (*e.g. CDC = Centers for Disease Control*). The UNDECIDED annotation helped to identify and subsequently include the few instances of these categories that occur in the *House MD* series in the acronym *Search* list.

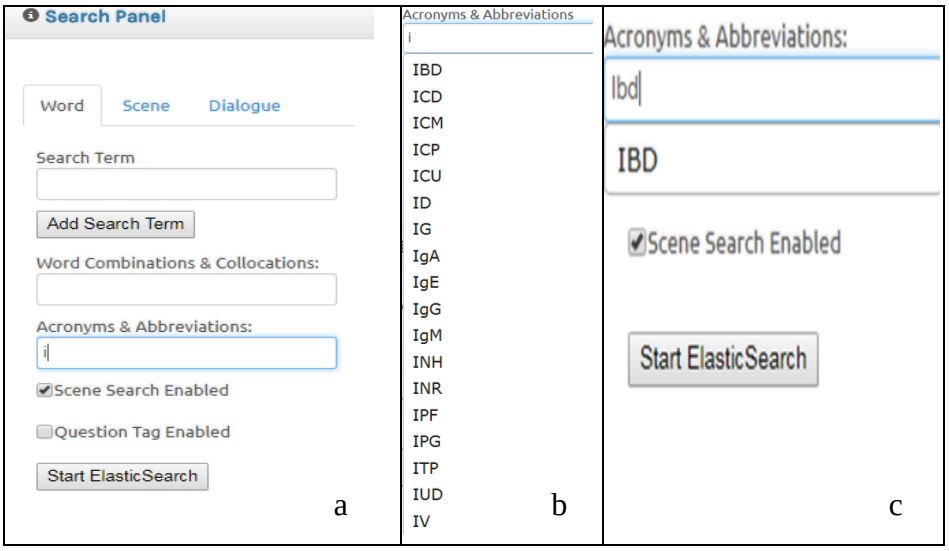


Figure 4.
Acronym searching: (a) list activation; (b) list restriction (c) item selection and search.

In this respect, as well as validating the acronym model, the work of annotation has also vindicated the choice of a TV series as a resource through which to engage with clinical acronyms. Certainly, medical opinion has long been divided on the clinical validity of TV healthcare – some supportive (e.g. Gordon *et al*. 1998), others more critical (Smith *et al*. 1972). However, the simulated hospital environment of the *House MD* series uses a large number of valuable clinical acronyms. Moreover, this TV series also prioritises clinical acronyms over other types of medical acronyms which is the reverse of what happens with many online medical acronym finders (e.g. *Acronym Finder: https://www.acronymfinder.com*) that prioritise healthcare personnel and administration acronyms over those relating to the diagnostic and therapeutic procedures that arise in a clinical context. As Figure 4 shows, the types of clinical acronyms found in the *House Corpus* are reassuringly those with which medical students need to engage.

The great care that the student annotators took needs to be mentioned. It was expected that undergraduate students in their early years of a language degree would make mistakes as regards the expansion of acronyms to their full

forms. There were, in fact, very few such cases. However, so far, neither the list of multi-word sources of acronyms (i.e. their full forms), nor the distribution across the corpus of the grammatical and functional properties of acronyms identified have been included in the *Search Panel* options of the *House Corpus* for reasons further explained in *Section 4*.

The second stage in the research consisted in the conversion of the 177 *Summary tables* thus created by the student annotators into a single table. From this, a *Search List* of acronyms was created that has been incorporated into the *Acronym Search* functionality, recently restyled as the *Acronym and Abbreviations* functionality, that is available in the *Search Panel* in the *House Corpus* interface. Figure 4 reconstructs the drop-down *Search Menu* used to make selections from this *Search List*; when a letter is typed into the search box, a list of acronyms starting with the corresponding letter appears; the typing of further letters, usually no more than two or three, further reduces the list until only one option remains, which can then be selected. Figure 5 shows how the *Search Result* functionality reports the number of search 'hits' for the query presented in Figure 4 (in this case two in the same scene).

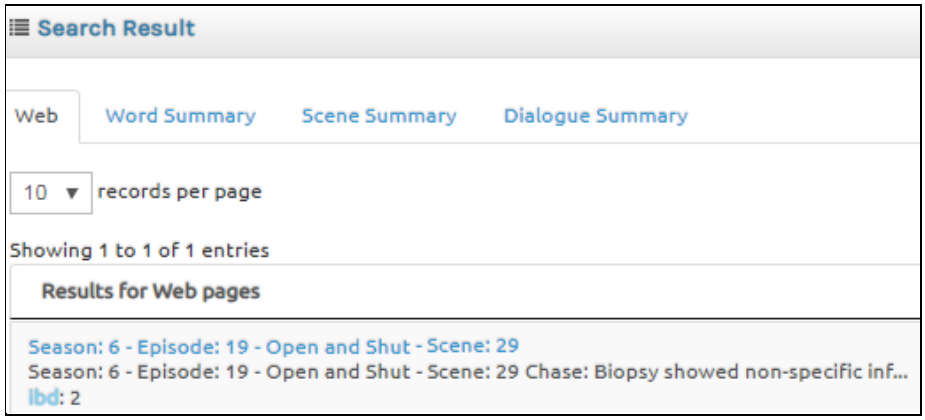


Figure 5
An example of the *Result Pages* of an acronym search.

Figure 6, instead, shows that clicking the hyperlink for a specific scene, Scene 29 in the example shown in Figure 5, ensures the written transcript is presented, accompanied by a scene viewing, with all instances of the searched-for acronym(s) highlighted. As many examples in this article show, the search possibilities include combinations with other words or acronyms. For example, selecting *ANA* from the acronym *Search List* and typing in *anti-DNA* in the *Word Combination & Collocations* box (Figure 4) returns a scene (not shown) where *anti-DNA* a.k.a. *anti-double stranded DNA* (*Anti-dsDNA*) *antibodies* are exemplified as a subgroup of *anti-nuclear antibodies (ANA).*

SEASON: 6 - Episode: 19 - Open and Shut - Scene: 29

CHASE: Biopsy showed non-specific inflammation, which suggests **IBD**, only it doesn't explain the cardiac problems.
TAUB: Or her declining kidney function. We ran panels during surgery.
HOUSE: Actually, **IBD** does make sense, 'cause it could trigger ankylosing spondylitis, which would explain the heart and the kidneys. Start her on sulfasalazine and **TNF** inhibitors. (They follow him in the corridor. House enters the elevator.) Oh, and Taub, I'm not gonna see you before you head off for Gomorrah. Just remember, if you get disoriented... breathe through your nose and look for the horizon.

Figure 6
An example of an acronym contextualised in a scene.

3. Results

Manual annotation revealed 324 different forms identified as medical acronyms (i.e. *types*) with a total of just under 3000 instances (i.e. *tokens*), a somewhat smaller figure than originally predicted. However, the online resource produced more occurrences than manual annotation as searches included the many examples in episode titles and stage directions and descriptions that students were instructed not to include in their annotation.

		Tokens	
Acronym	Full name	Manual	Online
5ASA	5-aminosalicylic Acid	2	2
B12	Cobalamin	1	14
B19	Buccal Neuron 19	1	3
BRCA1	Breast Cancer Gene 1	1	1
CA125	Cancer Antigen 125	2	2
CA19.5	Cancer Antigen 19.5	1	1
CD68	Cluster of Differentiation 68	1	1
CO2	Carbon Dioxide	1	1
FIO2	Fraction of Inspired Oxygen	1	1
Hba1c	Glycated Haemoglobin	2	1
MDR-1	Multi-Drug Resistance 1	1	1
NF2	Neurofibromatosis Type 2	8	8
O2	Oxygen	11	32
O2 sats	Oxygen Saturation	5	6
T2	Time for 63% of Transverse Relaxation	1	2
T3	Triiodothyronine (thyroid hormone)	1	1
T4	Thyroxine (thyroid hormone)]	3	3
		43	80

Table 3
Manual *vs.* online results for alphanumeric acronyms.

The higher number of tokens in online results was also due to more special circumstances. A small but significant percentage of clinical acronyms are alphanumerical, not all of them annotated owing to the insufficient initial

instructions given to the student annotators. Consequently, as Table 3 shows, for the 17 alphanumerical types identified, 43 tokens were annotated manually whereas 80 are returned by the online system.

Table 4 analyses acronyms' relative frequency in the corpus in terms of six frequency-related categories. As may be deduced from this, the vast majority relate to acronyms for which there is just one token in the corpus.

Category	Instances	Manual	Online
1	1	132	131
2	2-5	114	109
3	6-10	33	32
4	11-20	20	20
5	21-100	20	27
6	100+	5	5
Total		324	324

Table 4
Acronym Frequency per category.

Frequency is a crucial characteristic that the planned *Acronym Map* resource will take into consideration. Frequent acronyms pose different problems for students as compared with those that are less frequent. Thus, as stated above, *MRI*, the most frequent acronym in the corpus, can function as an adjective, noun or verb and can appear in inflected forms including prefixes (e.g. *fMRI*) or suffixes (e.g. *MRIs*) and can refer to a facility, procedure and as an adjective in multi-word combinations. *MRI* can also indicate test results or, collectively, refer to those who turn an MRI facility into a service. By contrast, at the time of writing, the entry in the online OED (Third edition) refers to *MRI* as a noun, but not to other parts of speech, and limits its definition to a medical procedure and associated equipment. Certainly, the OED's description of *fMRI* (*functional magnetic resonance imaging*) as a noun and adjective used to describe a medical procedure and the associated scan it produces is more comprehensive. The line *fMRIs tell us where the blood flow is* (*Scene 13, Episode 8, Season 5*) certainly indicates that the same inflectional processes that occur with *MRI* also occur with *fMRI* but none of the scenes where *fMRI* is referred to in the *House Corpus* illustrate the metonymic processes that have affected *MRI*. Inevitably, the planned *Acronym Map* resource will need to incorporate other sources that illustrate just how far these processes extend to *fMRI* in English and Italian discourse.

On both manual and online counts, over three-quarters of the acronyms occur only five times or less in the corpus. Many of these, terms like *IBD* (*inflammatory bowel disease*), occur in just *one* scene raising the question as to whether students should be required to learn terms that only appear once. As it happens, the IBD acronym was part of a survey of twenty frequent gastrointestinal acronyms sent to all medical house staff and attending

physicians in New York with a request to provide the full forms. This survey led the researchers to conclude that, "awareness of medical acronyms was less than acceptable" (Parakh *et al.* 2011, p. 9) since, gastroenterologists excluded, many of those asked were unable to give the correct reply. Such experiments clearly point to the need for acronym training and suggest that the thorough learning of all 300 or so acronyms present in the *House Corpus* constitutes a good investment for medical undergraduates.

Despite there being no *must-be-learnt* list of English medical acronyms for students in their pre-clerkship bioscience years, other ways of validating the *House Corpus* as a source of essential acronyms exist, one of which is to compare it with expectations about acronyms for the *USMLE Step 1* exam (www.usmle.org/step-1/), which assesses the first steps in medical studies. "Constructed according to an integrated content outline that organizes basic science material along two dimensions: system and process" it, alas, presents no to-be-learnt list of acronyms. However, the many *what-to-expect-in-USMLE-Step-1* primers available come close to doing so, as they contain lists of 'Common abbreviations' needed to pass the exam, which thus constitute a useful benchmark when evaluating acronyms for medical trainees. There are sufficient correspondences between *USMLE Step 1* and the levels of knowledge required of Italian students in their first years of studying Medicine to conclude that validating our acronym list in this way works.

T2	**TID**	**TPP thiamine pyrophosphate**
T3	**TIPS**	**TPR total peripheral resistance**
T4	**TM**	**TRH thyrotropin-releasing hormone**
TAL	**TMS**	**tRNA transfer ribonucleic acid**
TB	**TNF**	**TSH thyroid-stimulating hormone**
TBI	**tPA**	**TTP thrombotic thrombocytopenic purpura**
T-cell	**TPP**	
TEE	**TRH**	**TXA_2 thromboxane A_2**
THC	**TSH**	
TIA	**TTP**	
TIBC		

Figure 7
Side-by-side comparison of acronyms in the House Corpus and an USMLE Step 1 primer.

Thus the left-hand side of Figure 7 shows the list of 21 acronyms for the letter T in the *House Corpus*. The right-hand side of Figure 5 instead shows the seven 'Common abbreviations' for the letter *T* (all of them acronyms) from one such primer (Reinheimer 2005, p. xviii). Of the latter, four also occur in the *House Corpus* (TPP, TRH, TSH, TTP), suggesting that, although, as mentioned above, some integrations from other sources may be needed, the acronyms in our list do stand up to scrutiny.

We may conclude this Section by underscoring the fact that an *Acronym and Abbreviations* resource has been created that allows students to learn

acronyms in context. It can be used in conjunction with online tests (e.g. implemented through *Google Forms*) to encourage individual use in self-learning activities. Further improvements are planned such as enabling users to switch between an *acronym-only* version (e.g. *MRI*) and a version that includes reference to the multi-word source (i.e. stating that *MRI* refers to *Magnetic Resonance Imaging*). Others, such as the highlighting of *all* the acronyms in a specific scene, red for those searched-for but blue for the others, have already been undertaken as illustrated, for example, in Figures 3 and 6. However, the implementation of *Acronym Maps* is still some way off.

4. Discussion

As stated in the *Introduction*, clinical acronyms must be learnt in the early years of medical training. Figure 8, taken from a US website (http://tmedweb.tulane.edu/portal/student-guide/item/medical-terminology-and-abbreviations?category_id=20) with "the mission of providing our student community a website that brings together various facets of medical school", neatly summarises the reasons why medical students in their pre-clerkship years should invest in learning abbreviations.

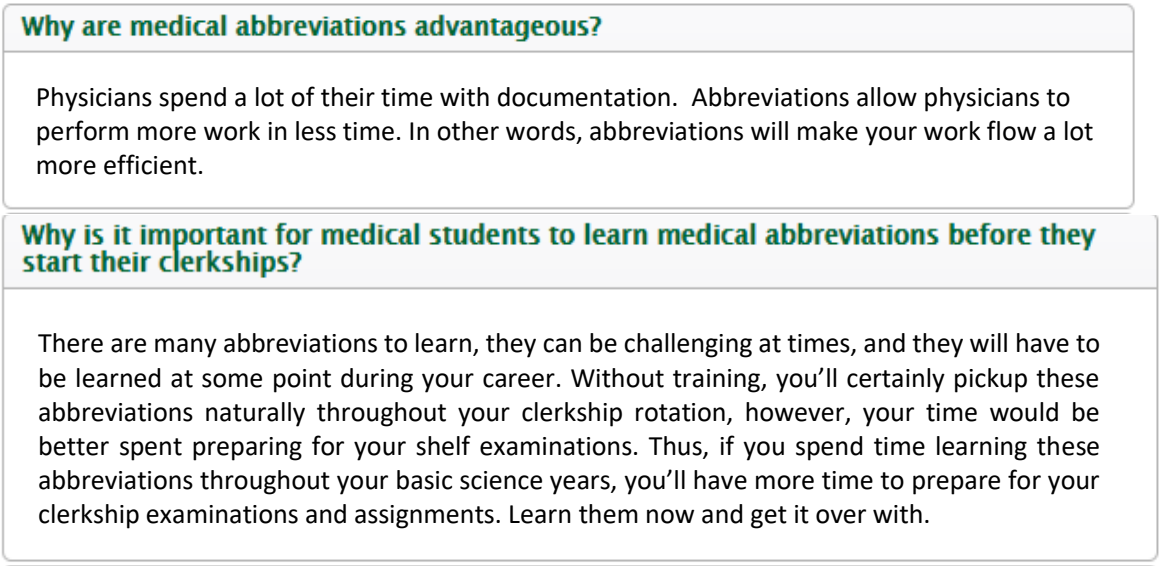


Figure 8
Learning medical abbreviations within a preparatory philosophy.

In an Italian context, the one for which the *Acronyms and Abbreviations* resource in the *House Corpus* is being developed, clerkships or rotations also occur. University administration describes them as *AFP Attività formative professionalizzanti* but they are referred to informally as *tirocini pre-laurea* and the students who participate in them as *tirocinanti.* As in other medical training

systems, *AFPs* are the first taste medical trainees have of working in a hospital setting.

As Figure 8 explains, after learning basic biomedical science, and as they approach the midway point in their degree course, medical trainees will spend an increasing amount of time learning the ropes in hospitals, 'rotating' through different medical specialties under the guidance and supervision of hospital doctors. In so doing, they learn how to treat and interact with patients by taking patient histories, carrying out physical examinations, completing questionnaires, writing up progress notes and taking their first steps in clinical training by watching what their supervisors and other hospital staff do. As the *House MD* TV series simulates many of these activities, the *House Corpus* with its scene-based structure (Taibi *et al.*, this volume) is potentially a good way to present the preparatory training advocated in Figure 8, whether in classroom lectures or online self-learning activities. Simulation characterises much medical training (Loiacono 2018, pp. 246-252) and, as in the case in question, provides an empirical basis on which a theoretical framework can be mapped.

For Italian medical students, however, there are other reasons why these acronyms must be learnt and taught, in particular in the context of the compulsory courses in English that medical students follow. Italian is one of the world's languages, which translates English medical acronyms less often than others (Gavioli 2005, pp. 92-94; Laviosa 2017, p. 20). Thus, while Italian typically uses the acronym *BPCO* (*Broncopneumopatia Cronica Ostruttiva*) which corresponds to *COPD* (*Chronic Obstructive Lung Disease*), it also refers to the *GOLD* (*Global Initiative for Chronic Obstructive Lung Disease*) guidelines for its assessment and the compound *GOLDCOPD* which appears in the official Italian website for the disease: *http://goldcopd.it*. Likewise, whereas most Romance languages use *SIDA*, formed from the initial letters of the full expression that has been used to translate this syndrome from English into many Romance languages (such as French, Spanish, Portuguese and Romanian), Italian, instead, adopts the English acronym: *AIDS.* There are, of course, many cases where Italian will avoid the use of an acronym altogether, preferring to talk about *immunodeficienza* – in other words, resorting to the use of part of the multi-word source as a very different but useful abbreviatory strategy.

Cos'è l'esame PSA? Il PSA (dall'inglese Prostate-Specific Antigen, ossia antigene prostatico specifico) è una proteina prodotta dalle cellule della ghiandola prostatica. L'esame ne misura i livelli nel sangue.
https://www.farmacoecura.it/tumore/valori-psa/
Il **PSA -** acronimo di ***Prostate Specific Antigen***, italianizzato in Antigene Prostatico Specifico - è una proteina sintetizzata dalle cellule della prostata. Piccole concentrazioni di antigene prostatico sono normalmente presenti nel siero di tutti gli uomini e si possono valutare tramite un semplice esame del sangue.
http://www.my-personaltrainer.it/salute/psa.html

Figure 9
Glossing the acronym PSA.

As Figure 9 shows, many online glossaries for the Italian general public exist that 'convert' English medical acronyms into their corresponding full forms in both Italian *and* English. So why bother about teaching acronyms? Are online glossaries not enough? In answer to these questions, the example shown in Figure 9 has not been chosen by accident. Reportedly, in the university where the author of this Section currently teaches, a senior academic asked a candidate to explain the meaning of this acronym during the medical student's final exam. Receiving no answer, the academic had to explain its meaning to the student. Even if a single instance is judged not to be sufficient justification for specific training of clinical acronyms, the authors' experience is that this is not the only example. What counts is the cumulative effect. Besides this, there are, in any case, other justifications relating to the need to transcend the lexicogrammatical aspects of acronyms and contemplate their discourse and digital aspects, described in the *Introduction*, that glossaries and other tools (such as those mentioned in the *tmedweb.tulane.edu* portal) do not – and probably cannot – contemplate.

What is of interest to teachers whose professional duties are to research and teach medical discourse in English is the potential of a specialised corpus based on scripted clinical discourse to illustrate the genre-related and sociolinguistic characteristics of acronyms, an important aspect of what, for want of a better label, may be termed 'clinical interaction theory'. Indeed, in the course of their clerkships, medical trainees will encounter medical, and above all clinical genres, many of which need to be understood and practised. The different uses to which acronyms are put are closely tied to specific medical genres. Within the preparatory and anticipatory learning context envisaged above, the corpus-based approach outlined in the previous sections seems to be a good solution for the contextualised learning of specific clinical acronyms, where 'contextualised' underscores their genre-related nature. This is the step that the planned *Acronym Maps* needs and intends to undertake.

Ironically, and somewhat paradoxically, it is precisely the confusion that surrounds the use of acronyms – the acronym soup often wittily served up in medical literature (Walling 2001, p. 14) – that constitutes a sound basis for persuading students to consider the status of medical acronyms in English as a discourse and genre-related problem rather than as a language problem.

Comparing examples from clinical manuals allows students to focus on the functions of clinical acronyms and not just on the forms they take. The text shown in Figure 10 is a passage from a volume on Emergency Medicine (Jenkins, Braen 2005, p. 6) for which an Italian translation has been published (Braen 2015, p. 4) and from which the text in the bottom part of Figure 10 has been taken.

One must remember that 1 to 2 minutes is required for medications administered at a peripheral site to reach the heart; this is true even when CPR is adequate. Most authorities therefore recommend that drugs be administered by rapid bolus and followed by a 20-mL bolus of fluid. When venous access is unobtainable, the following medications can be administered by endotracheal tube: lidocaine, epinephrine, atropine, and narcan (LEAN), which are administered in approximately 2- to 2.5-times the recommended dose, first diluted in 10 mL of normal saline, and injected by passing a catheter beyond the tip of the endotracheal tube. After injecting the medication, 3 to 4 forceful ventilations are provided.
• Occorre ricordare che sono necessari da 1 a 2 minuti affinché un farmaco somministrato in una zona periferica raggiunga il cuore anche nel caso in cui la CPR sia adeguata. • I farmaci vanno somministrati in bolo rapido seguito da un bolo di 20 ml di liquido. • Quando l'accesso venoso non è ottenibile i seguenti farmaci possono essere somministrati attraverso il tubo ET: lidocaina, adrenalina, atropina e naxolone LEAN (Lidocaine, Epinephrine, Atropine, Naxolone) somministrandone la dose consigliata in circa 2-2,5 volte, dapprima diluita in 10 ml di soluzione salina e quindi iniettata introducendo un catetere oltre l'estremità del tubo ET. • Dopo aver iniettato il farmaco si effettuano 3-4 ventilazioni forzate.

Figure 10
Glossing acronyms.

In keeping with the need for efficiency in Medicine described in Figure 8, the Italian text in Figure 10 underscores the need for straightforwardness in this medical specialty, in particular the need to give clear directions. It uses abbreviatory devices – bulleted presentation; omission of superfluous details such as "most authorities therefore recommend"; reduction of the English term *endotracheal tube* to the form *tubo ET* (where *ET* stands for *tubo endotracheale* i.e. *endotracheal tube*) – that shorten and sharpen the original text. All this is in addition to the abbreviatory devices used in the English text, as exemplified in both texts in Figure 10 in relation to the *CPR* procedure, a classic example of an acronym borrowed from English and used throughout Italian society in all healthcare-related services.

Note, however, the use of the term *LEAN* in Figure 10 both in the English and Italian texts to refer to a *group* of different entities in contrast to the common assumption that acronyms refer to a single entity. Indeed, leaving aside the difference in the English and Italian interpretation of N (*narcan* is the trade name; *naxolone* the name of the molecule), the term LEAN deserves a closer look. In *research articles*, it is described as an *acronym* (e.g. De Luca 2011, p. 681), but as a *mnemonic* in *manuals* (e.g. Davies, Hassell 2007, p. 14) and *handbooks* (e.g. Hughes, Mardell 2009, p. 462) and as a *mnemonic acronym* in *dissertations* (Bortle 2010, p. 158) – a demonstration, if ever one was needed, of the genre-based use and interpretation of medical acronyms.

How otherwise can the different name given to the very same term be explained, in particular when the authors took great in categorising their use of the LEAN abbreviation? The answer to the conundrum – *when is an acronym not an acronym?* – lies, of course, in the different uses to which it is put. In the case of the LEAN abbreviation, it is interpreted as a mnemonic in handbooks and manuals mindful of *don't-forget-to* clinical procedures but as an acronym where reflection on entities, as in research articles, predominates. In different clinical contexts and in different genres, mnemonics, like acronyms, undergo different degrees of formal and informal recognition and authorisation and hence transformation in their use, which students need to be made aware of. Every acronym has in theory the potential to become a mnemonic, and every mnemonic has the potential to become an acronym. Most will not exploit this potential, but some will, so that students need to be advised to look on the definitions given in dictionaries, such as those from the OED reproduced in Figure 9, not as watertight categories but as starting points in need of further refinement.

Acronym orig. *U.S.* **1.** A group of initial letters used as an abbreviation for a name or expression, each letter or part being pronounced separately; an initialism (such as *ATM*, *TLS*). 2. A word formed from the initial letters of other words or (occasionally) from the initial parts of syllables taken from other words, the whole being pronounced as a single word (such as *NATO*, *RADA*).
Mnemonic *n.* [..] 2 A device to aid the memory; (in later use) *spec.* a pattern of letters, ideas, or associations which assists in remembering something.

Figure 9
The online OED's definition of acronym and mnemonic.

Table 5 is what the author of this Section presents to her students as a way of underscoring the need to consider abbreviatory devices, regardless of whether they are formally known as acronyms, mnemonics, acrostics or something else, in terms of the actual functions they carry out, and, in particular, in relation to clinical genres and to those – doctors, healthcare workers, patients, researchers and others – who take part in clinical discourse. When medical acronyms are explicitly linked to the genres they enact, it becomes far easier to subcategorise their various forms. Most obviously, Table 5 makes use of functional labels that distinguish clearly between an entity and a procedure. This equips students with a device – the question probe – to decide whether in a particular clinical context a form is used to abbreviate (i.e. as an acronym) or to recall (i.e. as a mnemonic), in keeping with the dictionary definitions of these terms shown in Figure 9.

	Self-discourse	Doctor/HCW-patient discourse	Clinical Team/Trial discourse	Public Health discourse
PROCEDURE	Personalised Mnemonics	Questionnaire/Report Mnemonics	Checklist Mnemonics	Protocol Mnemonics
ENTITY	Personalised Acronyms	Medical Reports & Notes	Research Article/Clinical Trial Acronyms	Protocol Acronyms

Table 5
Acronym categories in clinical discourse.

Table 5 is also a starting point for the yet-to-be completed incorporation of the grammatical and functional tags of all the acronyms in the *House MD* series, already accomplished by student annotators and described above in *Sections 1* and *2* and which will be part of the planned *Acronym Maps* functionality. Details of how this can be achieved will need further discussion that takes various issues and considerable experimentation into consideration. One such possibility relates to incorporating online question probes in the exploratory form of a drop-down list of questions of the type: *What examples of Checklist Mnemonics exist in the House Corpus?* or *Does a Checklist Mnemonic ever appear in Doctor-Patient discourse in the House Corpus?* or even *Do acronyms used to describe Body States occur more frequently in Doctor-Patient discourse or in the discourse between Dr. House and his team*? As well as producing specific answers in the form of corpus 'hits', such probes can also help students appreciate the need for context to be taken into consideration and the need to reflect on the specific medical genres in which they are likely to occur. All this helps trainee doctors appreciate that, when experienced medical writers raise concerns in their discussions about the use of acronyms and mnemonics, their arguments are undermined when no reference to the genre(s) in which they are being used is made.

How do *question probes* link up with the categories described in Table 5 and with the idea of creating an *Acronym Map* functionality? In this respect, one such question probe might be *Is the term HIV an abbreviation for an entity or a mnemonic for a procedure that needs to be undertaken?* On the basis of the definition shown in Figure 9, the answer is, of course, that, as a pathological condition, it is an entity. However, when the same question is applied to the term *ABC* it is clear that the latter is a mnemonic as it fulfils the basic characteristic of all mnemonics in their role as a memory device. That is, mnemonics make explicit reference to an internalised checklist, listing the individual items to be performed in a procedure in a specific order. A mnemonic invites the user to pick out, perform and mentally tick off each item before proceeding to the next on the list.

The text in Figure 10 reconstructs this procedural aspect in a way that makes the ABC mnemonic's untrustworthiness explicit. Alas the writers of this text misleadingly describe the term 'ABC' both as an acronym and as a

slogan rather than as a mnemonic (current author's underlining). When they used the term 'slogan', they were, in fact, just one step away from providing a genre-referencing term of the type presented in Table 5. As illustrated below, most of the terms used in Table 5 *do* appear in the medical literature. Indeed, had the text in Figure 10 used the term *slogan mnemonic(s)* in criticising Public Health campaign slogans, it would have connected up with other instances and made its authors' arguments more powerful. Indeed, the criticism of the false reassurances that Public Health *slogan mnemonics* generate is not confined to the text in Figure 10. It resurfaces in other healthcare texts (see Loiacono 2018, *Chapter 11* for *Médecins sans frontières*' criticisms of the slogans used in the UN and WHO's promotion of the SDG and MDG programmes).

> Today's most commonly cited acronym for HIV prevention – "ABC" – falls severely short of describing the global effort needed to reduce HIV transmission. First, the ABCs mix up different prevention strategies. "A" (for abstinence) and "B" (for be faithful) are behaviors. "C" (for condoms) is a commodity. The implication of this string of concepts is that anyone can achieve protection if he or she chooses one or more options from the short menu. [….] The "alphabet soup" approach overlooks interventions needed to protect people in risk-filled environments such as prisons or refugee camps. The ABCs infantilize prevention, oversimplifying what should be an ongoing, strategic approach to reducing incidence. True, the simplicity of the ABC slogan has probably helped some people better appreciate that they can take basic steps to protect themselves from HIV infection. But that advantage must be weighed against the dangerously misleading messages the ABCs send to both individuals and to policy makers. "ABC" gives the incorrect impression that all HIV transmission is sexual and that effective prevention is simply a matter of changing the individual choices of millions of people with a few, tried and true interventions. Reciting The "ABCs" invites distracting and useless arguments, such as whether abstinence is better than partner reduction or both are better than condom use […] The alphabet soup approach ignores core components of a comprehensive prevention response and the critical importance of adapting programming to distinct epidemics. Key aspects of prevention programming are invisible in the ABCs. In Eastern Europe nearly two thirds (62%) of new HIV infections reported in 2006 were due to non-sterile injection drug use. (Collins et al. 2008:)

Figure 10
Acronyms and dangerously misleading messages.

As the text shown in Figure 10 is not addressed to *intra-hospital* clinical care, it will not respond to the question probe – *is this abbreviation a clinical entity or a clinical procedure?* It nevertheless represents a useful starting point as regards the need to go beyond mere knowledge sharing as it considers trust and reassurance in medical discourse as significant in all aspects of medical discourse, an aspect in which the *House MD* TV series excels.

The categories established in Table 5 need to be briefly described. Though traditionally labelled as an acrostic, the term *personal memory device* (or PMD) used in Table 5 seems more appropriate as it is essentially a way of checking that nothing has been left out in the answers given in clinical exams. Unlike other categories, they are personal and not intended to be shared with others, though many successful doctors are keen to hand down to students the PMDs they themselves invented as trustworthy devices to pass exams in their student days. The example in Figure 11, with the PMD shown in brackets and 'indexed' with an icon, is from Reinheimer (2005, p. 16).

What three muscles constitute the erector spinae?	1. **I**liocostalis 2. **L**ongissimus 3. **S**pinalis ("**I L**ove **S**cience" muscles) 

Figure 11
Mnemonics Medical Trainees use as checklists in exams.

While personalised abbreviations are not designed to be shared, the evidence from online clinical blogs, forums and associated threads shows that they *are* assumed to exist.

For Goodness Sake - stop using personal acronyms

This was the transfer note for a patient. Please, if you use acronyms like these, just stop! Take the extra second and type it out.

Following SBAR format report given to Charge Nurse SBAR report
S- Situation (describe the condition of patient): Pt was admitted for Upper GIB, received 2 units of PRBC and 1 L of NS in MICU. HCt 19->23. Hemodynamically stable. BP 120-130s/60s, SR 70s-80s. Satting 98-100% on RA. LS clear t/o. Abd snt,BT x4. BM today, dark green/melena soft formed medium. Edema to BL LE +3 L>R. PPP+. BG before lunch, SS as ordered, MD ordered.
B- Background (concise/pertinent history of patient): See problem list and ICU/Pulmonary notes.
A- Assessment (your conclusion of patients condition now): Pt is hemodynamically stable. HCT. No c/o syncopy, SOB, CP, n/v/d. Able to transfer from bed to commode w/ stand by assit only and ambulate for short distance in room.
R- Recommendation (what needs to be done for/with the patient when they get to new location?): Transfuse 2 units PRBC as ordered at 1200. Please discusss with MD, higher insulin coverage for noon CS, SS 5 units administered as ordered. Monitor for s/sx of bleeding. CS check q AC &HS.

Comment thread [selected items]

- Acronyms like which? On first read I don't see any that are new to me.
- As I was reading it I kept saying to myself what personal acronyms? Then I got down to the comments and saw this was the consensus thoughout. That made me pretty happy because I have only been an RN for 4 months so I thought maybe I am missing something.
- Glad to know I'm not the only one without a problem reading this...
- What acronyms are bugging you? This read very easily to me. The acronyms used here are used at both hospitals I work at?

https://www.reddit.com/r/nursing/comments/1z8boj/for_goodness_sake_stop_using_personal_acronyms

Figure 12
Personal acronyms: how standard are they?

As the text in Figure 12 demonstrates, this awareness of their existence often surfaces in medical discourse where the boundary between shared and unshared gets blurred typically where outsiders complain about 'personal acronyms' that insiders consider as shared conventions.

The example and *Comment Thread* reproduced in Figure 12 makes it clear that in clinical practice agreements about what can be used and what cannot be used are based on consensus and experience rather than on formal agreements. The text, some parts of which have been omitted but which has

otherwise been reproduced in its original form, implicitly illustrates SBAR's transition from mnemonic to genre status. The *Comment Thread* section shows readers have no trouble with the acronyms used because of their familiarity with, and experience of, the *SBAR format report* (i.e. a genre). SBAR belongs to the second category in Table 5, the one in which doctors and other healthcare workers (HCWs) write reports about individual patients, sometimes as a result of questionnaire-based interactions with their patients. While outsiders or trainees will, of course, have difficulty with this genre, it needs to be recalled that *SBAR* is one of the commonest written 'mnemonic' genres, so well known that it has influenced the development of oral mnemonics such as I-PASS designed to prevent miscommunication in handovers (Starmer *et al.* 2012, p. 201).

Consensus is thus a vital aspect of acronym use. What *can* and *cannot* be used has been established by the *JOINT COMMISSION* considered by many to be the final arbiter. Its website (www.jointcommission.org/) explains that apart from those on a short list of unacceptable abbreviations, any reasonable standardization of abbreviations, acronyms, and symbols is acceptable and also holds (third bullet point in Figure 13) that personal acronyms are by no means automatically disbarred.

Abbreviation List – Options

Is a list of acceptable abbreviations required? No. The requirements found at IM.02.02.01 do not require organizations to maintain a list of acceptable abbreviations. Developing and maintaining a list of acceptable abbreviations would be an organizational decision. IM.02.02.01 EP 2 requires that organizations use 'standardized' abbreviations. Any reasonable approach to standardizing abbreviations, acronyms, and symbols is acceptable. Examples may include:

- Standardized abbreviations developed by the individual organization.
- Use of a published reference source. However, if multiple abbreviations, symbols or acronyms are used for the same term, the organization identifies what will be used to eliminate any ambiguity.
- A decision that individuals who work in the organization may use any abbreviation, acronym, or symbol that is not on the list of unacceptable abbreviations. However, if multiple abbreviations, symbols, or acronyms exist for the same term, the organization identifies what will be used to eliminate ambiguity.

Figure 13
Personal acronyms: how standard are they?

We may note in passing that their control on naming processes is far less than that carried out by *The United States Adopted Names Council* (*USANC:* www.ama-assn.org/about/united-states-adopted-names-council). This latter agency approves generic names for drugs, and hence abbreviations, in the US (Loiacono 2013b, pp. 31-32). On the contrary, besides inviting users to suggest acronyms to be added to their list, the *FDA (US Food and Drug Administration*) goes no further than providing a list of them stating that:

> "The emphasis is on scientific, regulatory, government agency, and computer application terms. The database includes some FDA organizational and program acronyms".
> *www.fda.gov/aboutfda/fdaacronymsabbreviations/default.htm*

The transition to clerkship is inevitably a moment of truth when cultural assumptions about acronyms and mnemonics come to be scrutinised. Mnemonics and acronyms that are part of a clinician's PMDs have the habit of slipping out and causing consternation and surprise. A question probe of the type suggested above might take the following form: *Are personal abbreviations ever used or contested in clinical contexts?* This would lead to the scene in *House MD* shown in Figure 14 (one of seven scenes in this series where mnemonics are discussed) and would function at the very least as a basis for further discussion about the sociomedical functions (disruptive or constructive?) of personalised uses of abbreviatory devices.

SEASON: 3 - Episode: 13 - Needle in a Haystack - Scene: 10

FOREMAN: (puts up x-ray) Take a look at this. Bones of your forearm.
STEVIE: Uh. Radius and ulna.
FOREMAN: How about the wrist?
STEVIE: Um. Lunate. Hamate, The...
CAMERON: Scared lovers try positions they can't handle. (Foreman and Stevie look at her) It's a **mnemonic** for the wrist bones. It's the only way I can remember them.
STEVIE: Ow.

Figure 14
Personalised mnemonics.

A medical trainee's first real taste of English abbreviations used in Italian clinical care are those indicated in the second column of Table 5. They arise in the context of questionnaires used in doctor-patient interviews. Although the term *questionnaire* appears just once in the *House Corpus*, student encounters with patients are well represented and include encounters with elderly patients, among the most frequent types of patient interview that medical trainees perform. *Monitoraggio e valutazione delle ACTIVITIES OF DAILY LIVING* (www.tlc.dii.univpm.it/blog/wpcontent/uploads/2014/10/ADLs_per_Sito.pdf), is the part English, part Italian title of an online teaching document by Susanna Spinsante, Università Politecnica delle Marche, illustrating the questionnaires that students use in these encounters. The document describes them in Italian but with constant reference to partly translated, partly untranslated documentation in English. We will not explore the various aspects of these questionnaires – *ADL (Activities of Daily Living), IADL (Instrumental Activities of Daily Living), MMSE (Mini Mental State Examination)* – that test a patient's autonomy cognitively and physically, except to further characterise the second column in Table 5. In this context, the medical trainee looks on these terms as entities and not as procedures. As the answer to each step is recorded, the trainee is guided by the printed or electronic questionnaire as regards the steps to be undertaken. While the term *questionnaire acronyms* is well established in the

medical literature (Cleemput, Dobbels 2007; Forsén *et al.* 2007), the author of this Section has yet to find a convincing example in the medical literature of the term *questionnaire mnemonics* during such interviews, a matter that has generated considerable discussion with students.

The contrary is true with the next category (Column 3 in Table 5) which indicates the explicit recognition in the medical literature of *checklist mnemonics* as exemplified by the text in Figure 15. This text recognises the value of a *mnemonic checklist* (current author's underlining) in stressful clinical contexts, and the usefulness of a specific mnemonic in reducing clinician error and promoting awareness among medical students.

> Metacognition is a cognitive debiasing strategy that clinicians can use to deliberately detach themselves from the immediate context of a clinical decision, which allows them to reflect upon the thinking process. However, cognitive debiasing strategies are often most needed when the clinician cannot afford the time to use them. A mnemonic checklist known as TWED (T = threat, W = what else, E = evidence and D = dispositional factors) was recently created to facilitate metacognition. This study explores the hypothesis that the TWED checklist improves the ability of medical students to make better clinical decision […] it is a predecided measure that allows the automatisation of goal intentions even in unfavourable environments (e.g. a busy and stressful environment). For example, if the intended goal is to minimise diagnostic errors secondary to cognitive biases, the implementation intention could be the use of a mnemonic checklist, like the TWED checklist, which is memorable and easily retrievable.

Figure 15
TWED a Mnemonic checklist Chew *et al.* 2016, pp. 694-697).

Often invented and shared by specific clinical teams, *checklist mnemonics* are a preventive measure countering tiredness, encouraging focus and detachment from the distractions in a hospital environment and stimulating teamwork. It is thus hardly surprising that checking lists is a major part of medical training and practice, a matter constantly foregrounded in the *House MD* series and indeed many other TV medical soaps.

It is never too early to make students aware of the power of abbreviations to persuade (Loiacono 2013a) which includes the downsides of an abbreviatory device like a *mnemonic checklist* which attribute enticing and even amusing names to clinical trials but which can sometimes hide more sinister realities. The sociolinguistic and psycholinguistic aspects of acronyms in clinical practice, specifically in randomized trials, are highlighted in the text in Figure 16, (current author's underlining). As well as establishing the continuities between acronyms and mnemonics and between research findings and clinical practice, this text also points out that, while randomised trials certainly come to end, the acronyms and mnemonics they use may not. In so doing, the text suggests the more subtle and insidious uses that such inventions may subsequently perform.

….. these results support the hypothesis that naming randomized trials with an acronym may enhance the citation rate. This is consistent with the function of acronyms in human language as effective mnemonic tools. Their influence might also be subliminal, since specific acronyms could invoke subconscious value-laden associations that might enhance positive perceptions of the studies they name, a phenomenon in cognitive psychology known as "automatic attitude activation." Enhanced attention to and recall of studies through the use of acronyms may facilitate the appropriate translation of research findings into clinical practice. If acronyms exert influence independently of normative markers of clinical credibility, however, such influence is not rational scientifically, even if it is understandable psychologically. Consequently, this subtle linguistic tool could undermine evidence-based practice. The observed close association between acronym use and sponsorship by the pharmaceutical industry amplifies this concern.

Figure 16
(Stanbrook, Redelmeier 2006, pp. 101-102: subliminal functions of acronyms and mnemonics).

The final category (Column 4 in Table 5) relates to the passage of acronyms and mnemonics from the status of abbreviations used by a specific clinical team in a specific clinical trial to that of a protocol, a much higher status that, once again, has its pros and cons. This is the stage where the socially shared status of *mnemonic checklists* and *mnemonic acronyms* shifts from clinical experimentation to a more universal level of recognition, in part thanks to the prior consensus achieved. Thus, as the SIGHT abbreviation illustrates, medical abbreviations for protocols typically undergo a staged process of approval and assessment: they are first recommended or strongly advised (Figure 17), then made compulsory (Figure 18) and finally proposed as candidate for international protocols (Figure 19).

Key recommendations

1. Clinicians (doctors and nurses) should apply the following mnemonic protocol (SIGHT) when managing suspected potentially infectious diarrhoea:

S	Suspect that a case may be infective where there is no clear alternative cause for diarrhoea
I	Isolate the patient and consult with the infection control team (ICT) while determining the cause of the diarrhoea
G	Gloves and aprons must be used for all contacts with the patient and their environment
H	Hand washing with soap and water should be carried out before and after each contact with the patient and the patient's environment
T	Test the stool for toxin, by sending a specimen immediately

Figure 17
The mnemonic protocol for CDI: NHS England 2013.

SIGHT was coined because of the marked increase in outbreaks of CDI (*Clostridium difficile* infection), attributable to the (mis)use of antibiotics, which led to a European surveillance protocol (*https://ecdc.europa.eu/sites/portal/files/documents/European-surveillance-clostridium-difficile-v2point3-FINAL_PDF3.pdf*). The explicit recognition of abbreviations like SIGHT as a *mnemonic protocol* in the medical literature (Figures 17 and 18) is conditioned by many factors so that the change in status is a gradual process, the result of constant negotiation. As Figure 17 shows, the SIGHT abbreviation is a UK

national protocol; its interpretation by specific NHS Trusts, such as the *Solent NHS Trust* (www.solent.nhs.uk/), through detailed letter-by-letter analysis, has extended its influence. Besides the basic recommendation, Figure 18 reproduces the part of SIGHT relating to the letter G.

(p.6) Clinical staff must apply the following mnemonic protocol (SIGHT) when managing suspected potentially infectious diarrhoea.
(p.7) Gloves and Aprons (Personal Protective Equipment) • On entering the room, staff must wash hands with soap and water and wear an apron and gloves. • Visitors who do not assist in patient care and who have minimal patient contact do not need to wear gloves and an apron. • Visitors assisting with patient care should wear gloves and an apron. • All visitors and staff should wash their hands with soap and water before they leave the room. • Visitors or staff should not eat or drink in the vicinity of the patient. • On leaving the room all staff or visitors (who wear gloves and aprons) must remove and dispose of apron and gloves into the clinical waste bin and wash hands using soap and water
http://www.solent.nhs.uk/_store/documents/ipc11policyforthepreventionandcontrolofclostridiumdifficileinfection.pdf May 2015

Figure 18
Clinical implementation of a mnemonic protocol for CDI: NHS Solent 2015.

On the contrary, the text in Figure 19 (Wiuff *et al*. 2018, p. 15) hints at the difficulties in approving protocol mnemonics beyond national borders, which suggests that the fight against antibiotic resistance first needs to tackle resistance to the use of English as a *lingua franca*.

> When discussing European practice for CDI treatment, variability between countries is inevitable for a number of reasons. Treatment of patients with CD begins with making diagnosis, specifically having a high index of clinical suspicion if a patient has a combination of signs and symptoms and/or CDI risk factors and thereafter conformation by microbiological testing or colonoscopic/histopathological findings. Clinician awareness of CDI as part of the differential diagnostics is therefore crucial for appropriate patient management. However, there remains considerable variability across countries with an estimated 40,000 inpatients potentially undiagnosed annually in European hospitals […]. Mnemonic checklists can be useful tools to reduce clinician error and promote awareness […]. Albeit potentially more useful when English is the commonly spoken language, the SIGHT mnemonic is a useful aide memoire for clinicians when managing patients with suspected potentially infectious diarrhoea …

Figure 19
Negotiating European-wide protocol status: the SIGHT mnemonic for suspected CDI.

Somewhat surprisingly, no mnemonic is used by the Italian CDI protocol which instead provides a summary of actions to be implemented (*Schema riassuntivo azioni da implementare*: http://internetsfn.asl-rme.it/cio/pdf/Protocolli/ 201014 _clostridium_difficile_rev0_14.pdf). Despite the fact that the *g* of Italian *guanti* coincides with the *g* of English *gloves*, the details in the list differ and include face masks as well as gloves and aprons. Leaving to one side issues of whether abbreviations like SIGHT are in best interests of European citizens, the real task facing teachers of medical English is describing

and promoting descriptive frameworks that allow a detailed comparison of the process of summarising and abbreviating within protocols.

This leads us back to the issue of how to instil awareness of different abbreviatory solutions to similar problems in different languages and cultures and the need to explore ways in which corpora can provide the necessary detail on which to pin such comparisons. One candidate for this role is the transformation of *House M.D.* episodes into *clinical vignettes*, a special type of clinical teaching case used primarily to measure trainees' knowledge and clinical reasoning. Essentially a medical vignette describes a hypothetical patient's age, gender, medical complaint and health history (Converse *et al.* 2016, p. 588) using a stepped procedure, as explained in the *I-TECH Clinical Mentoring Toolkit* document entitled *Structured Clinical Vignettes: What Are They and How Are They Used?*:

> Vignettes are structured according to the classic sections of the medical visit—chief complaint; history; physical exam; laboratory and radiographic studies; assessment and plan—presented in chronological order to the trainee. Each section consists of a narrative describing the situation, followed by a question or series of questions prompting the trainee to explain how she or he would care for the patient, given the information presented. The trainee indicates what she or he would do, not by selecting from a fixed list of multiple choice options, but by providing a detailed explanation of steps. This requires trainees to apply their knowledge to the situation, much like […] in an actual patient visit. *www.go2itech.org/HTML/CM08/toolkit/tools/vignettes.html*

Although vignettes are used in exams to encourage analysis of a specific diagnosis or clinical situation or to measure trainees' skills in performing the tasks necessary to diagnose and care for a patient (Nendaz *et al.* 2000; Scalese, Hatala 2013; Holmes, Ponte 2011), the process can be harnessed to test students writing skills, i.e. summarising the reasoning and skills displayed in a TV medical drama in the form of a vignette. Insofar as they present patient-related cases and scenarios involving unusual diseases and unusual presentations of common diseases with an educational value, the episodes in *House M.D.* mimic clinical vignettes and provide a useful framework when encouraging the proposed summarising. Asking students to consider why they think, for example, that CDI is discussed without using the acronym form (*SEASON: 6 - Episode: 05 - Instant Karma - Scene: 04*) in contrast to the use of MRSA (*SEASON: 6 - Episode: 18 - Knight Fall - Scene: 13*) could be the basis of a student's reconstruction of a clinical vignette relating to hospital-acquired infections that summarises these two episodes in a structured way. Although researchers often discuss the issue of abstract writing in ESP and medical training (Dudley-Evans 2002; Griffin, Hindocha 2011) as a desirable vocational skill, undergraduate students in their pre-clerkship years do not have the research experience to achieve this. On the contrary, writing a summary of

a TV episode in keeping with the clinical vignette framework would appear to be a better first step as it provides practice in the art of clinical writing that includes learning how to abbreviate.

To sum up: the research and the support we have received from student annotators has helped promote an understanding of the Pinocchio-like process of conversion of strings of letters into the lifelines that international protocols constitute, but also the snares – some exaggerated, others genuine – on the road to consensus in the use of abbreviations. Acronyms have a life of their own and are not pseudo-words. Whether they present themselves as entities or procedures, they can easily change their forms and functions; they can be borrowed and loaned between languages and genres and can be avoided completely or alternatively invented to give new meanings to existing words, often in a way that is designed to amuse, tantalize and tease. In this sense, they are a continuation in contemporary Medicine of a long line of genres and literary devices that explore amusing ambiguities and paradoxes in word formation – puns, analogies, limericks, metaphors presented as riddles, enigmas and conundrums – many of which can be traced back to the earliest days of English literature (Loiacono 2012) whose origins lie in what has been described as "conscious semantic exploitation" (Pons-Sanz 2014, p. 24).

Perhaps more importantly, the above discussion has established a distinction between: genres that use acronyms and mnemonics in the clinical context (*physical examinations, patient interviews* and *associated questionnaires*); genres that talk about their use in the clinical context (*research articles, handbooks, manuals, dissertations*); genres purely for training and assessment purposes (*primers, clinical vignettes, medical textbooks*). In so doing we have merely scratched the surface as regards a genre-related approach to the learning of abbreviatory processes in medical discourse. Only a brief mention has been made above, for example, of the use of acronyms and mnemonics in handovers (a.k.a. handoffs) and the communication hurdles that have to be overcome, succinctly but safely, when one clinical team (e.g. the 'day' shift) is replaced by another (e.g. the 'night' shift). Nor have we discussed other reflections on acronyms made in medical research genres, such as review articles which, in order to provide state-of-the-art assessments, summarise and weigh up findings about specific topics published in the medical literature and which presuppose a capacity to reconcile different abbreviatory forms and strategies. The fact that at least one review article exists dealing specifically with the 'handoff mnemonics literature' and which reviews '46 articles describing 24 handoff mnemonics' (Riesenberg *et al.* 2009, p. 196; see also Mardis *et al*. 2016) is a clear demonstration of the need to extend what has so far been achieved with the *House Corpus Acronym and Abbreviations* resource.

The need to contemplate different *categories* of medical genres has been underscored many times above. The provision of *Acronyms Maps* suggests one way in which this might be done. Most of the categories mentioned in Table 5 are likely to be represented in the day-to-day work of clinical activities, such as differential diagnosis, whereas those genres used in training and assessment are more likely to include a higher proportion of abbreviations relating to the first two columns in Table 5. In the early stages of medical education, this is probably enough. While a clear boon for medical English classrooms, such maps may also support hunches about differences in the nature and incidence of acronyms in spoken and written forms of medical discourse in English as well as differences with other languages, e.g. Italian, whose oral medical discourse would seem to place less reliance on acronyms than English does. Generally speaking, the more *Acronym Maps* can be retrieved from specialised corpora, such as the *House Corpus*, the better, as this may well encourage greater consideration in corpus studies of specialised genres and contexts. In the case of spoken medical discourse, such studies seem to be particularly urgent (Loiacono 2016).

5. Conclusions

Learning to abbreviate is an essential part of learning how to communicate in any profession, as it requires good judgements to be made. A fine balance has to be achieved in medical communication between clarity of meaning and compact expression. Training medical students, regardless of whether English is their first language or not, to master the use of abbreviatory devices in medical discourse in English, requires clearly-defined descriptive models that illustrate the process of abbreviation at work, ones that, where appropriate, take the different practices of medical discourse in different languages, such as Italian, into account.

Terms like ‘acronym’ and ‘mnemonic’ relate to many different realities that need to be explained to medical trainees in their first years of medical education. Yet, despite medical journals’ heavy investment in online learning, a recent search into online archives such as *The BMJ* and *NEJM* revealed little in support of the learning of abbreviatory processes. The *Acronym Search* resource that the authors have developed for the *House Corpus* is a much-needed first step in this direction. By familiarising students with the realities of acronyms in clinical care in their pre-clerkship years, an awareness has been created of the pitfalls that medical writers have signalled (Baue 2002; Brubaker, Brubaker 1999; Cheng 2003; Kuhn 2007; Patel, Rashid 2009; Pottegård *et al.* 2014; Summers, Kaminski 2004). However, more importantly, a significant step has been made as regards encouraging students to compare

abbreviatory processes in different languages such as English and Italian and to make their *own* judgements about when, and when not, to abbreviate, which includes an awareness of the impracticalities, and in many cases the absurdity, of the demands in the medical literature for acronyms to be abolished or curtailed.

It will be clear from what has been stated above that specialised corpora are needed to satisfy general educational requirements in Medicine. The *House Corpus* shows that the role of specialised corpora can go beyond a mere support for the learning of specific acronyms, promoting instead an awareness of descriptive rather than prescriptive models of the use of abbreviations in clinical care. However, if descriptive approaches are to win the day over prescriptive ones that have muddled thinking and which merely tend to confuse medical students, then a better link-up between medical systems and medical genres is required (Loiacono 2012). In the current project, further work is already underway to fulfil the requirement for the *House Corpus* to incorporate genre-related searches in its interface. A greater focus on the abbreviating process is justified and might be achieved, for example, by encouraging students to 'convert' episodes in the *House MD* series into clinical vignettes.

Addressing the issue of how representative the acronyms included in the *House Corpus* are with regard to those which students meet in their early years of medical training (see *Section 4*) requires further research and assessment. Evaluation of a corpus and its search functionalities is never easy, owing to the co-presence of mutually confounding factors. We are comforted, in this respect, by the insights expressed by others who have used acronyms in their corpus research studies in view of their expectation for "technical acronyms to be relatively stable across languages" (Baroni, Bernardini 2004, p. 1313). We are also reassured by the fact that benchmarking is possible and is indeed a quality-assessment exercise that has a long tradition in corpus studies and in the development of online e-learning resources in Higher Education. In a world of uncertainties, providing medical students with reassurances about the right road to take in their studies of medical discourse is both demanding and at the same time a source of considerable satisfaction. The more research draws on the reassuring footing of corpus linguistics, the more it shines light on the need for further research to be undertaken into the process of abbreviation, whose role in medical communication is all too frequently underestimated and misunderstood.

Acknowledgements: Francesca Bianchi (Università del Salento) is thanked for co-ordinating the work of student annotators without which this research would not have been possible.

Bionotes: **Anna Loiacono** is Associate Professor in the University of Bari's Department of Basic Medical Science, Neuroscience and Sense Organs. She previously held a similar position in the University of Foggia. With vast experience in teaching, publishing and promoting Medical English training projects in the biomedical sector, her publications include *The Medical Alphabet Vol. 1* (2013) and *The Medical Alphabet Vol. 2* (2018), Andria: Matarrese Editore.

Francesca Tursi is a CEL specialising in Medical English, working at the University of Foggia's Language Centre.

Authors' addresses: anna.loiacono1@uniba; francesca.tursi@unifg.it.

References

Baroni M. and Bernardini S. 2004, *BootCaT: Bootstrapping Corpora and Terms from the Web*, in *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*, pp. 1313-1316.

Baue A.E. 2002, *It's acronymania all over again: with due reference to YB Yogi Berra*, in "Archives of Surgery" 137 [4], pp. 486-489.

Berlin L. 2013, *TAC: AOITROMJA? (the acronym conundrum: advancing or impeding the readability of medical journal articles?*), in "Radiology" 266 [2], pp. 383-387.

Bortle C.D. 2010, *The role of mnemonic acronyms in clinical emergency medicine: A grounded theory study*, Doctoral dissertation, University of Phoenix.

Brubaker R.F. and Brubaker J.H. 1999, *Does somebody else out there hate acronyms*?, in "Archives of Ophthalmology*"* 117 [5], pp. 701-702.

Cheng T.O. 2003, *Please let every acronym be defined (PLEAD*), in "Catheterization and cardiovascular interventions" 60 [3], pp. 424-425.

Chew K.S. Durning S.J. and van Merriënboer J.J. 2016, *Teaching metacognition in clinical decision-making using a novel mnemonic checklist: an exploratory study*, in "Singapore Medical Journal" 57 [12], pp. 694-700.

Cleemput I. and Dobbels F. 2007. *Measuring patient-reported outcomes in solid organ transplant recipients*, in "Pharmacoeconomics" 25 [4], pp.269-286.

Collins C., Coates T.J. and Curran J. 2008, *Moving beyond of the alphabet soup of HIV prevention*, in "AIDS" 22 [Suppl 2], pp. S5-S8.

Converse L., Barrett K., Rich E. and Reschovsky J. 2015, *Methods of observing variations in physicians' decisions: the opportunities of clinical vignettes*, in "Journal of General Internal Medicine" 30 [3], pp. 586-594.

Davies J.H. and Hassell L.L. 2007, *Children in Intensive Care: A Survival Guide*, Elsevier, China.

De Luca D., Cogo P., Zecc E., Piastra M., Pietrini D., Tridente A., Conti G., and Carnielli V.P. 2011, *Intrapulmonary drug administration in neonatal and paediatric critical care: a comprehensive review*, in "European Respiratory Journal" 37 [3], pp. 678-689.

Dudley-Evans T. 1997, *Genre models for the teaching of academic writing to second language speakers: Advantages and disadvantages*, in Miller T. (ed.), *Functional Approaches to Written Text: Classroom Applications*, United States Information Agency, Washington, pp.150-158.

Federiuk C.S. 1999, *The effect of abbreviations on MEDLINE searching*, in "Academic Emergency Medicine" 6 [4], pp. 292-296.

Forsén L., Loland N.W., Vuillemin A., Chinapaw M.J., van Poppel M.N., Mokkink L.B., van Mechelen W. and Terwee C.B. 2010, *Self-administered physical activity questionnaires for the elderly*, in "Sports Medicine" 40 [7], pp.601-623.

Gaudan S., Kirsch H., and Rebholz-Schuhmann D. 2005, *Resolving abbreviations to their senses in Medline*, in "Bioinformatics" 21 [18], pp. 3658-3664.

Gault L.V., Shultz M. and Davies K.J. 2002, *Variations in Medical Subject Headings (MeSH) mapping: from the natural language of patron terms to the controlled vocabulary of mapped lists*, in "Journal of the Medical Library Association" 90 [2], pp. 173-180.

Gavioli L. 2005, *Exploring corpora for ESP learning*, John Benjamins, Amsterdam/Philadelphia.

Gordon P.N., Williamson S. and Lawler P.G. 1998, *As seen on TV: observational study of cardiopulmonary resuscitation in British television medical dramas*, in "BMJ: British Medical Journal" 317 [7161], pp. 780-783.

Griffin M.F. and Hindocha S. 2011, *Publication practices of medical students at British medical schools: experience, attitudes and barriers to publish*, in "Medical Teacher" 33 [1], pp. e1-e8.

Holmes S.M. and Ponte M. 2011, *En-case-ing the patient: disciplining uncertainty in medical student patient presentations*, in "Culture, Medicine, and Psychiatry" 35 [2], pp.163-182.

Hughes S. and Mardell A. (eds.) 2009, *Oxford handbook of perioperative practice*, Oxford University Press, Oxford.

Jenkins J.L. and Braen G.R. (eds.) 2005, *Manual of emergency medicine*, Lippincott Williams & Wilkins, Philadelphia; Italian trans.: Braen G.R. 2015, *Manuale di Medicina D'Emergenza*, Delfino, Roma.

Kuhn I.F. 2007, *Abbreviations and acronyms in healthcare: when shorter isn't sweeter*, in "Pediatric nursing" 33 [5], pp. 392-401.

Laviosa S. 2017, *Empirical translation studies: from theory to practice and back again*, in Pagano A., Laviosa S., Kemppanen H. and Ji M. (eds.), *Textual and Contextual Analysis in Empirical Translation Studies*, Springer, Singapore, pp. 1-26.

Loiacono A. 2012, *Excelsa Grammatica: Dal Piers Plowman a Pearl*, Cacucci Editore, Bari.

Loiacono A. 2012, *Medical communication: Systems and genres*, Ibis, Como-Pavia.

Loiacono A. 2013a, *Sociomedical interaction in English: towards virtual hospitals*, in "JAHR" 4 [7], pp. 195-215.

Loiacono A. 2013b, *The Medical Alphabet: An English Textbook in Healthcare, Vol.1*, Matarrese, Adria.

Loiacono A. 2016, *Forward Revisited: English-Language texts and films on emergency medicine*, Matarrese, Adria.

Loiacono A. 2018, *The Medical Alphabet: An English Textbook on Healthcare in the Digital Age Vol 2. G-M.*, Matarrese, Adria.

Mardis T., Mardis M., Davis J., Justice E.M., Holdinsky S.R., Donnelly J. and Riesenberg L.A. 2016, *Bedside shift-to-shift handoffs: a systematic review of the literature*, in "Journal of nursing care quality" 31 [1], pp. 54-60.

Nendaz M.R., Raetzo M.A., Junod A.F. and Vu N.V. 2000, *Teaching diagnostic skills: clinical vignettes or chief complaints?*, in "Advances in health sciences education" 5 [1], pp.3-10.

Pakhomov S. 2002, July, *Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts*, in *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, pp. 160-167.

Parakh P., Hindy P. and Fruchter G. 2011, *Are we speaking the same language?: acronyms in gastroenterology*, in "The American Journal of Gastroenterology" 106 [1], pp.8-9.

Patel C.B. and Rashid R.M. 2009, *Averting the proliferation of acronymophilia in dermatology: Effectively avoiding ADCOMSUBORDCOMPHIBSPAC*, in "Journal of the American Academy of Dermatology" 60 [2], pp. 340-344.

Pons-Sanz S.M. 2014, *The language of Early English literature: from Cædmon to Milton*, Palgrave Macmillan, New York.

Pottegård A., Haastrup M.B., Stage T.B., Hansen M.R., Larsen K.S., Meegaard P.M. and Aagaard L. 2014, *SearCh for humourIstic and Extravagant acroNyms and Thoroughly*

Inappropriate names For Important Clinical trials (SCIENTIFIC), in "BMJ*"* 349, g7092.

Reinheimer B.A. 2005, *USMLE Step 1 Recall: Buzzwords for the Boards*, Lippincott Williams & Wilkins, Philadelphia.

Riesenberg L.A., Leitzsch J. and Little B.W. 2009, *Systematic review of handoff mnemonics literature*, in "American Journal of Medical Quality" 24 [3], pp. 196-204.

Scalese R.J. and Hatala R. 2013, *Competency assessment*, in Levine A., DeMaria Jr. S., Schwartz A.D., Sim A.J. (eds.), *The comprehensive textbook of healthcare simulation*, Springer, New York, pp. 135-160.

Schuemie M.J., Kors J.A. and Mons B. 2005, *Word sense disambiguation in the biomedical domain: an overview*, in "Journal of Computational Biology" 12 [5], pp. 554-565.

Shultz M. 2006, *Mapping of medical acronyms and initialisms to Medical Subject Headings (MeSH) across selected systems*, in "Journal of the Medical Library Association" 94 [4], pp. 410-414.

Smith F.A., Trivax G., Zuehlke D.A., Lowinger P. and Nghiem T.L. 1972, *Health information during a week of television*, in "New England Journal of Medicine" 286 [10], pp. 516-520.

Stanbrook M.B., Austin P.C. and Redelmeier D.A. 2006, *Acronym-named randomized trials in medicine—the ART in medicine study*, in "New England Journal of Medicine" 355 [1], pp. 101-102.

Starmer A.J., Spector N.D., Srivastava R., Allen A.D., Landrigan C.P., Sectish T.C. and I-PASS study group 2012, *I-pass, a mnemonic to standardize verbal handoffs*, in "Pediatrics" 129 [2], pp. 201-204.

Stevenson M., Guo Y., Al Amri A. and Gaizauskas R. 2009, *Disambiguation of biomedical abbreviations*, in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, Association for Computational Linguistics, pp. 71-79.

Summers J.B. and Kaminski J. 2004, *Acronym addiction*, in "Texas Heart Institute Journal" 31 [1], pp. 108-109.

Walling H. 2001, *When will the MEK inherit the ERK*? *Acronym alphabet soup*, in "Trends in pharmacological sciences" 22 [1], p. 14.

Wiuff C., Banks A.L., Fitzpatrick F. and Cottom L. 2018, *The Need for European Surveillance of CDI*, in Mastrantonio P. and Rupnik M. (eds.), *Updates on Clostridium difficile in Europe*, Springer, Cham, Switzerland, pp. 13-25.

Xu Y., Wang Z., Lei Y., Zhao Y. and Xue Y. 2009, *MBA: a literature mining system for extracting biomedical abbreviations*, in "BMC bioinformatics" 10 [14], pp 1-10.

DISAMBIGUATING NEAR SYNONYMS IN MEDICAL DISCOURSE
A multilayered corpus analysis of *disease*, *illness* and *sickness* in the British National Corpus

STEFANIA M. MACI, RÉKA R. JABLONKAI, MAREK ŁUKASIK, SOPHIKO DARASELIA, DANIEL KNUCHEL[‡]

Abstract - This paper discusses the preliminary results of a corpus-based analysis of three basic health-related lexical items: *disease*, *illness* and *sickness* on the British National Corpus (henceforth BNC) CQP Web platform (2007 XML). Synonymous at first glance, the terms exhibit a certain degree of co-text and context semantic variation; therefore, the lexical items in question cannot be used interchangeably. This in turn may pose some difficulties in inter-lingual translation and language learning, mainly stemming from the lack of full equivalence (or, in some instances, zero equivalence) between the words and their counterparts in some other languages, such as German or Italian. The paper aims to demonstrate how collocational behaviour and semantic profiles can help disambiguate near synonyms along a cline between general and specialised discourse. To this end, the study employed corpus linguistic methods and analysed the BNC across all its text genres. The collocational patterns of the three selected lexical items were examined in the corpus and the semantic profiles of the lexical items were established. The findings suggest that the three health-related near synonyms exhibit markedly different collocational behaviours and semantic preferences. It is therefore suggested that the approach adopted in this study could be applied to help disambiguate the meanings of near synonyms appearing in any specialised discourse at both intra- and inter-linguistic levels. Future research will compare the findings resulting from a similar investigation to be carried out on COCA to see the extent to which, if any, (a) meanings can vary and (b) whether meaning variations associated with these items depend on the interactants (i.e. professionals/laymen).

Keywords: medical discourse; corpus-based analysis; statistics; semantic preference; semantic profile; collocation.

[‡] The research is the result of a joint effort of the five authors. However, Daniel Knuchel and Marek Łukasik wrote the *Introduction*, *Method* and *Conclusions* sections; Réka Jablonkai wrote the *disease* paragraph the *Methods* and *Conclusions* sections; Stefania M. Maci revised the *Introduction*, *Methods*, *Conclusions* sections, and wrote the *sickness* paragraph and the *Conclusions* section; Sophiko Daraselia wrote the *Illness* paragraph and the *Conclusions* section. The investigation carried out by Stefania M. Maci is part of the MIUR-funded research project of national interest (PRIN) *Knowledge Dissemination across media in English: continuity and change in discourse strategies, ideologies, and epistemology* (COFIN grant no. Prot. 2015TJ8ZAS_002), directed by prof. Marina Bondi of the University of Modena and Reggio Emilia – local coordinator, prof. Maurizio Gotti.

Doi Code: 10.1285/i9788883051531p127

1. Introduction

The history of English medical discourse dates back to the 17th century, when the role of Latin diminished in favour of vernacular languages. This coincides with some important changes in the medical profession at the time, mainly the shift from a scholastic (authority and/or dogma-based) to an evidence-based approach (Taavitsainen 2018). Indeed, as Taavitsainen (2018, p. 252) explains, while genre conventions change slowly, some elements acquire new connotations. Evidence-based medicine and its approach mean the practice of medicine is hierarchically based on scientific evidence, with a crucial role attributed to "patient values and preferences in clinical decision making, and the development of the methodology for generating trustworthy recommendations" (Djulbegovic and Guyatt 2017, p. 415). In this context, there is a need for reliable and correct information, because humans 'consume' information, or better, "[h]umans are "informavores" - we need evidence to effectively function in the world around us" (Djulbegovic and Guyatt 2017, p. 421).

Also, new discoveries, novel procedures, methodologies and sophisticated equipment have led to an exponential increase in the numbers of terms and new medical genres, all of which have been necessary for the precision-oriented communication of professional knowledge (Gotti 2016).

At the beginning of the 20th century, the language of medical science around the world still used three languages to roughly equal extents: German, English and French (Baethge 2008). However, similar to many other specialised fields (Crystal 2002), English took the lead in the mid-20th century, particularly in scientific publications, and its dominance has prevailed ever since in worldwide professional medical communication (Baethge 2008). On the other hand, national languages are still widely used locally, in doctor-patient communication in individual countries, in teaching and in scientific activity; furthermore, the dominance of English on the Internet in non-English speaking countries has declined, as evidenced by Baethge (2008) and Graddol (2006, p. 14). The use of English in scientific publications and native languages on the Internet corroborates the presence of some form of linguo-pragmatic dichotomy, which is rooted in both systematic and linguistic differences (that still persist). For example, anatomical terms and the names of diseases are imported directly with their correct Latin endings into Germanic languages, such as German or Dutch, while they are more readily naturalised in Romance languages, such as French or Italian (Wulff 2004). In addition, languages can also borrow medical terms from English. Borrowings also occur between languages other than English (Wulff 2004). These processes result in a number of inconsistencies (at various levels) and paint a highly complicated picture of

current medical language, which might prove challenging for professional medical translators. At the same time, health communication contexts in English-speaking countries are becoming increasingly multilingual while English is often used as a lingua franca in doctor-patient communication or between medical professionals (Sentel and Braun 2012). As many doctors, nurses and other healthcare workers are non-native speakers of English, special attention should be paid to their use of English in healthcare contexts and it should be ensured that they use the right expressions and terms when communicating (Candlin and Candlin 2003). It is vital that misunderstandings or imprecise uses of language are avoided as these might complicate diagnoses and/or disease treatment.

The aim of this study is to explore how corpus linguistic methods can be applied to disambiguate the meaning of the selected health terms. This study is part of a larger-scale project that investigates the lexical behaviour and semantic profiles of some selected health terms in several other languages which belong to the same or different language families. We hope to gain insights into potential translation problems of medical terms and phrases from English into other languages, for instance, Georgian, German, Italian, Hungarian and Polish, and vice versa.

More specifically, this study aims to identify the core meanings of three near synonymous lexical items in English, namely: *disease*, *illness* and *sickness*. These basic health-related words are investigated through their collocations generated for the entire British National Corpus (henceforth BNC) CQP Web platform (2007 XML). Collocations constitute the basis on which meaning analyses can be carried out, given that they highlight the most frequent semantic fields within which each lexical item can be grouped.

2. Near synonyms and semantic preference

Our study is based on the assumption that meaning is a pragma-semantic construct (Wittgenstein 2003; Busse 2015, pp. 91-122). Therefore, although words may seem to be synonymous, they might differ in their use. The relationship between such words is often referred to as near synonymy in the literature. Xiao and McEnery (2006) define near synonyms as "lexical pairs that have very similar cognitive or denotational meanings, but which may differ in collocational or prosodic behaviour" (p. 108). Moreover, such near-synonymous words have specific semantic profiles which we understand as cognitive and denotational meanings, plus their use in context (which grants the speaker pragmatic knowledge). In other words, the semantic profile of a word is a broad meaning-driven sketch of this word. Especially in language learning, it is crucial for non-native speakers to have pragmatic knowledge of

L2 in general and to know the word sketches for near synonyms (Barron 2003, cf. also Baker *et al.* 2013). In particular, when there are no equivalents in L1 and L2, near synonyms are difficult to translate. An example is the German word *Krankheit*: although it can be translated into English as *disease*, *illness* or *sickness*, it is fundamental to distinguish their different semantic profiles, which will help users to use the words appropriately and express intended meanings precisely. Other examples include Italian words, such as *malattia*, which can be translated into English as *disease*, *illness* or *sickness*. As aptly underlined by Loiacono (2018, p. 398), the distinction between the terms *illness* and *disease* is an endogenous/ exogenous one, so that the term *illness* should refer to the state or condition of the disease, whereas the term *disease* refers to the type of disease itself. As a consequence, this suggests differences in frequencies of the singular/ plural forms of the two items. Yet, there are blurred cases due to external social forces, especially among laymen. The issue is not simply a linguistic problem: things are far more complex, because the distinction between *illnesses* and *diseases* is the distinction between patients and professionals:

> Patients suffer "illnesses", doctors diagnose and treat "diseases" […] when physicians dismiss illness because ascertainable "disease" is absent, they fail to meet their socially assigned responsibility. It is essential to reintegrate "scientific" and "social" concepts of disease and illness as a basis for a functional system of medical research and care. (Engberg 1977 in Loiacono 2018, p. 399)

A clear understanding of the semantic profiles of the English terms will facilitate the selection of the most appropriate equivalent in any given context.

To understand the semantic profile of a word we can look at "the company it keeps" (Firth 1957 [1951], p. 11). Firth introduced the term collocation to describe word associations and their impact on meaning. In the last six decades, several studies on collocation have been published and different empirical methods have been tested for the analysis of collocational patterns of lexical items (cf. Brezina *et al.* 2015; Gablasova *et al.* 2017). In this study, we align with Sinclair's (2004, p. 28) definition of a collocation as "a frequent co-occurrence of words".

Previous analyses of collocates of particular lexical items have found that lexical items have a tendency to co-occur with "other words that belong to a particular semantic set" (Hunston 1995, p. 137). Stubbs (2001), for example, defines semantic preference as "the relation, not between individual words, but between a lemma or word-form and a set of semantically related words" (p. 65). To illustrate this category of relation, Partington (2004) gives the semantic preferences of *sheer*. The semantic sets the word *sheer* was found to collocate with included (1) "magnitude", "weight" or "volume", e.g.

the sheer volume of reliable information; (2) "force", "strength" or "energy", e.g. *the sheer force of his presence*; (3) "persistence", e.g. *sometimes through sheer insistence*; (4) "strong emotion", e.g. the sheer joy of life and (5) physical quality, e.g. *the sheer glamour of evil* (p. 145). Furthermore, he demonstrated that there is also interaction between typical syntactic behaviours of words and their semantic preferences. In the example of *sheer*, the typical structure for the first two semantic sets, that is, "magnitude" and "force" words, was found to be "*the sheer* (noun phrase) *of* (noun phrase)". In the third semantic category, the word *sheer* was found to be often preceded by prepositions expressing means or manner, e.g. *through, out of, by*. Nelson (2006) found that, in business discourse for example, the word *package* had a preference for being connected to computers, and it also shared a preference related to finance, with words like *merger* and *market.*

Stubbs (1995) also demonstrates that lexical items have a tendency to co-occur with negative or positive words. This phenomenon is usually referred to as semantic prosody in the literature (Partington 2004). In his analysis of the word *cause*, for example, Stubbs (1995) found that its most frequent collocations are negative abstract nouns like *anxiety, concern* and *crisis,* and many examples are from the medical field, like *cancer, blood, death,* and *disease*. Furthermore, Nelson (2006) claims that looking at the semantic prosody of words as used in business discourse not only reveals insights into language use, but also provides information about the business world as such. The examples he gives here are semantic prosodies of the words *boss* and *manager*. According to his findings, *boss* has a tendency to be used with negative adjectives, such as *meanest* and *old-fashioned*, whereas *manager* displays positive collocates, like *excellent* and *good* (Nelson, 2006). In addition, Partington (2004) also analyses the relationship between semantic preference and semantic prosody. He suggests that in most cases semantic prosody can be considered a sub-category of semantic preference, a special case that includes "instances where a lexical item shows preference to co-occur with items that can be described as bad, unfavourable or unpleasant, or as good, favourable or pleasant" (p. 149). In a further analysis, however, he notes that "semantic preference is a 'narrower' phenomenon - relating the node to another item for a particular semantic set - than prosody which can affect wider stretches of text" (p. 151). He also illustrates how semantic preference contributes to building semantic prosody and how prosody in turn restricts the preferential choices of the node word. Several authors (Baker *et al*. 2008; Bednarek 2008), however, argue that a clear distinction should be made between semantic preference and semantic prosody. Bednarek (2008) proposes that, following Sinclair (2004), the term semantic preference should exclusively be applied to "collocations of a lexical item with (more or less specific) semantic subsets" (p. 132), and the term semantic prosody used for

all other attitudinal and evaluative meanings which often go beyond being merely positive and negative. The present study focuses on semantic preference when attempting to identify nuanced differences in meaning and usage between the selected health-related lexical items.

3. Methodology

3.1. Selection of lexical items

The terms selected for the analysis in the current study were chosen on the basis of two independent pilot analyses: (1) a comparison of dictionary definitions of the health-related lexical items, (2) synonym-finding query applied to the BNC (BYU-BNC at corpus.byu.edu, see Davies 2004).

3.1.1 Dictionary definitions

Definitions of the terms *disease*, *illness* and *sickness* were analysed as regards their synonyms in three online English dictionaries, namely, the Oxford dictionaries, including the Oxford English Dictionary, the Collins English Dictionary, and one of the most popular online medical dictionaries, i.e. the medical Merriam-Webster Dictionary (https://www.merriam-webster.com/medical). Although the investigation is based on the BNC, the fact that the medical Merriam-Webster Dictionary is American does not pose any problem. Indeed, the last attested time when the medical community stressed the linguistic importance of any difference existing between the British Medical Dictionary and the Webster American Dictionary was in 1962 (Talbott 1962), and ever since it has not been dealt with.

The results indicate that, overall, the selected lexical items are perceived in general language as being synonymous in relation to one another, with the reservations that (a) the study should be treated as approximate, insofar as the strength of synonymy relations is not provided in any of the aforementioned dictionaries and (b) in some cases one synonym is simultaneously offered as the *genus proximum* in the definition (i.e. in the *definiens* part), which results in a circular definition rather than an indication of a semantic position of the *genus* against the *definiendum*.

The term ‘disease’ was chosen as the point of departure, being the most generic and overarching medical term to represent the concept of interest, namely, that of “a disorder of structure or function in a human, animal, or plant, especially one that produces specific symptoms or that affects a specific location and is not simply a direct result of physical injury” (cf. https://en.oxforddictionaries.com/). Oxforddictionaries.com presents the

terms *illness* and *sick* as synonyms of *disease*. An 'illness' is "a disease or period of sickness affecting the body or mind"; sickness is "[t]he state of being ill" (https://en.oxforddictionaries.com/). In *Collins*, a 'disease' is an "illness which affects people, animals, or plants, for example one which is caused by bacteria or infection". An 'illness' is a particular disease, such as measles or pneumonia. *Sickness* is the state of being ill or unhealthy (https://www.collinsdictionary.com/).

The Merriam-Webster medical dictionary gives the following definition for 'disease': "an impairment of the normal state of the living animal or plant body or one of its parts that interrupts or modifies the performance of the vital functions, is typically manifested by distinguishing signs and symptoms, and is a response to environmental factors (as malnutrition, industrial hazards, or climate), to specific infective agents (such as worms, bacteria, or viruses), to inherent defects of the organism (as genetic anomalies), or to combinations of these factors: sickness, illness". The condition of having a *disease*, therefore, is that of having a combination of the two factors of sickness and illness. An 'illness' is defined as the "unhealthy condition of body or mind: sickness and 'sickness' as "the condition of being ill: ill health" or the condition of having a "specific disease" (https://www.merriam-webster.com/medical).

According to the OED, these words entered the English language and became part of the English lexicon at different times. In particular, the term 'sick' is a common Germanic word and is attested to in Old English (from 700 AD), whereas 'ill' is used in Early Middle English (from 1200 AD) and 'disease' (from 1300 AD). The words have developed senses that were associated with some of their first meanings and usage. For example, *ill* has been synonymous with *evil* (although not etymologically related) from the 12th century, which resulted in different meanings from 'sick' or 'disease'. For the purposes of this research, we will compare OED senses with those of the BNC corpus. In this analysis, we do not consider obsolete meanings as these do not add much information to this comparative study.

The word 'disease' originated in Middle English (1150 to 1500), meaning the "absence of ease, uneasiness, discomfort" (OED). In the OED, *disease* as a noun has three distinctive meanings:

1. Absence of ease; uneasiness, discomfort;
2. A condition of the body, or of some part or organ of the body, in which its functions are disturbed or deranged; Also applied to a disordered condition in plants;
    a. The condition of being out of health; illness, sickness;
    b. An individual case or instance of such a condition; an illness, ailment, malady, disorder;

   c. Any one of the various kinds of such conditions; a species of disorder or ailment, exhibiting special symptoms or affecting a special organ.
3. *Figurative*: A deranged, depraved, or morbid condition (of mind or disposition, of the affairs of a community, etc.; an evil affection or tendency.

The term *illness* derives from the adjective *ill*, introduced during the Middle English period with the sense of 'morally wicked'. Its meaning as a substantive refers to the quality or condition of being ill (in various senses). The OED records it with the meaning of "bad or unhealthy condition of the body (or, formerly, of some part of it); the condition of being ill; disease, ailment, sickness, malady".

The term 'sickness' derives from the adjective 'sick', with the sense of 'suffering from physical ailment', and was introduced during the Old English period. The substantive has four distinctive senses indicated by the OED as follows:

1. The state of being sick or ill; the condition of suffering from some malady; illness, ill-health (also figuratively);
2. A particular disease or malady (also in a figurative sense). It may also refer to a defect in wines or to a disease in sheep which can cause braxy;
3. A disturbance of the stomach manifesting itself in retching and vomiting;
4. Figuratively, it indicates utter disgust or weariness.

Although all dictionaries indicate that a *sickness* is a state or a temporary condition, they also suggest that the terms are not interchangeable. Yet they have not been useful in defining these words: *oxforddictionaries*, for instance, has indicated that *disease* is the overarching term, and that *illness* can have as synonyms both *disease* and *sickness*, while *sickness* can be a synonym only of *illness*. The OED reveals that *disease* has as synonyms both *illness* and *sickness*, and that synonyms of *illness* can be *disease* and *sickness*, but the synonym of *sickness* can only be *illness*. *Collins*, on the other hand, suggests that *disease* is an *illness* and vice versa, but that the condition of *sickness* is given by an *ill* state. The medical Webster-Merriam Dictionary indicates that a *disease* can be a *sickness* or an *illness*, that an *illness* is a *sickness* and that a *sickness* is a condition reflecting a *disease* in which one person is *ill*. Clearly, the use of dictionaries is not enough and this must be implemented with synonym-finding queries in the BNC.

3.1.2 Synonym-finding queries in the BNC

Synonyms were also found in the BNC via synonym-finding queries carried out through the corpus.byu.edu platform. The lemma rather than the word-form was chosen as a query node. The corresponding absolute frequencies of the three lemmas are as follows: DISEASE (f=8,799 singular; 1,817 plural), ILLNESS (f=3,194 singular; 506 plural) and SICKNESS (f=1,186 singular; 14 plural).

In a search of the whole corpus for synonyms of DISEASE, *illness* ranks first, with *sickness* coming eighth on the rank list, preceded by *disorder* (f=1,604), virus (f=1,474) and syndrome (f=1,197). In an analogous query for ILLNESS, *disease* comes first, with *sickness* ranking tenth, preceded, again, by *virus* and *syndrome*. The BNC (BYU version), apparently, does not help us to disambiguate these terms. For this reason, a more in-depth analysis is necessary.

3.2. The corpus

The research is based on queries applied to the entire British National Corpus through the University of Lancaster UCREL CQP Web platform[1] (see Hardie 2012). CQP Web offers access to the 2007 XML edition of the BNC, which comprises 112,102,325 word tokens and 638,862 word types, derived from 4,048 text samples. The corpus is POS-tagged using the BNC Basic Tagset (also known as C5),[2] and offers rich metadata, allowing the researcher to compare fine-grained sets of data across various categories (parameters). The simple search mode allows queries of the entire corpus or the written/ spoken mode only. The system also supports more advanced searches using Simple Query Syntax.

3.3. Methodological approach

With the aim of drafting semantic profiles of the three lexical items under investigation, we conducted a series of lexical analyses, mainly with the use of simple frequency counts (absolute and relative frequencies, AF and RF, respectively), dispersion measures (mean frequency, standard deviation and Juilland's D (distribution tests) and collocation extraction statistics (log likelihood and log ratio), as described below.

[1] https://cqpweb.lancs.ac.uk/ (17.3.2018).
[2] http://www.natcorp.ox.ac.uk/docs/c5spec.html (24.3.2018).

Dispersion measures are necessary in order to avoid biased results for AF and RF. Indeed, they helped us to compute frequencies of occurrence and co-occurrence of the three lemmas under investigation. Since in isolation statistical tests may be misleading, as each single test does not take into consideration the degree of dispersion, the three tests were used together to yield sound results (cf. also Gries 2008).

Collocates are words which usually go with the word under investigation (the node) and are computed within a range distance from the node of 5-1 words to the left and right of the node itself (Hunston 2002). In order to be significant, statistical tests, normally log-likelihood (LL) and log-ratio (LR) tests, determine their frequency (Hunston 2002; McEnery *et al.* 2006). Since one statistic for collocation extraction may yield skewed results, we applied both LL and LR in order to obtain more reliable data.

3.3.1. General data across the BNC

In order to gain an overview of the use of the three lexical items in the entire BNC as well as across specific genres, we applied both AF and RF, dispersion measures (mean frequency, standard deviation and Julliand's D) as well as a simple text count (Brezina *et al.* 2015). We also investigated the ratio between singular and plural forms of the three lexical items in the whole corpus. In all other studies lemma-based queries were used.

For easier handling of the data, we introduced text categories that differ from the original BNC text types and include the following: ACADEMIC, FICTION, NON-ACADEMIC, NEWS, OTHER WRITTEN and SPOKEN.

3.3.2. Collocation analysis methods

It was assumed that collocations at the textual level have the potential to reveal the semantic features of lexical items and their underlying concepts. Therefore, in order to determine the semantic profiles of DISEASE, ILLNESS and SICKNESS, we conducted six collocation queries, two for each term.

We employed a lemma-based noun-only collocation search, with the collocation window defined as 5L-5R and minimum collocate and node-collocate frequencies set to 5. We used two statistical measures to extract collocations, namely, log likelihood (or LL, employing significance statistics) and log ratio (or LR, measuring the effect size) (see e.g. Evert and Hardie 2014; McInnes 2004). While LL scores collocations by statistical significance, LR measures how big the difference is between the (relative) frequency of the collocate alongside the node, and its (relative) frequency in the rest of the corpus or sub-corpus. We therefore sorted collocates by LL and LR: the presence of a collocate in both lists was the inclusion criterion for

collocates in the present study. For the purpose of the semantic analysis, statistical significance was set as $p \leq 0.05$.

The next step after generating collocation lists for each statistical measure was to compare the results and extract the top 150 common collocates from both lists. The collocates were then analysed for semantic preference, taking a corpus-driven approach. Altogether, 23 semantic categories were identified in the study (see Table 1) and each collocate was assigned to one of them. The list was extended every time a new category emerged from the analysis. The final step included calculating the most frequent semantic categories for each term in order to draft their semantic profiles.

Semantic tag	**SEMANTIC CATEGORY**
TRANSMISSION	how it develops, contagious, how it is transmitted, inherited
NAME	specific name of a disease/illness/sickness
BODY PART	reference to a body part, location of the disease/illness/sickness in the body function of the body
EFFECT	reference to symptom, effect of the disease/illness/sickness
CAUSE	reference to cause of the disease/illness/sickness
LACK	reference to a lack of food
QUALITY	quality or characteristics of the disease/illness/sickness
TYPE	reference to the type of disease/illness/sickness
WHO/WHAT	reference to whom or what is ill
FUNCT. WD.	function word
GEOGRAPHY	geography
QUANTITY	quantity
TREATMENT	treatment
LEGAL	reference to a legal document, legislation
TIME	reference to age, time period in life
PREVENTION	prevention
SYNONYM (SG)	sg that is similar to disease/illness/sickness
DIFFERENT	different, various, other
EXAMPLE	example, such as, including
SPECIFICITY	specific, certain
INCIDENCE	incidence, case, one specific example
RISK	risk
ONSET	start, onset of a disease
INSTITUTION	institution
SUPERSTITION	popular belief, superstition

Table 1
Semantic categories identified for the most frequent collocates
of DISEASE, ILLNESS and SICKNESS in the BNC.

Having explained our methodological approach, we will now turn to the results of the data analysis in the next paragraphs.

4. Results and Discussion

4.1 General comparison of the collocational behaviour of the three lexical items

Table 2 below presents the overall frequencies of the three lexical items analysed in this study. Please consider that the different frequency results from those indicated in paragraph 3.2.1. are due to the fact that the BNC BYU is a different version of the BNC CQP Web platform, one which seems more complete and offers a wider range of options and apparently more reliable results (cf. also https://www.uni-bamberg.de/fileadmin/eng-ling/fs/Chapter_11/Index.html?3123ExerciseforBYUBNC.html [09/12/18]).

As can be seen from the summary in Table 2, there seems to be a strong preference for the use of *disease* rather than *illness* and *sickness*, and overall, the relative frequency is higher in the written sub-corpus than in the spoken one. There is also a marked preference for the use of lexical items in the singular form.

Lemma	No. of texts	Absolute Frequency (spoken)	Relative Frequency/million words (written/spoken)	singular/plural (%)
disease	1,214	10,680 (291)	95.27 (103.8/24.3)	83 / 17
illness	1,029	3,718 (214)	33.17 (35/17.9)	86 / 14
sickness	528	1,209 (101)	10.78 (11.7/8.4)	99 / 1

Table 2
Absolute and relative frequencies of *disease*, *illness* and *sickness* in the BNC.

As regards the types of texts under consideration, details are given in Table 3 below. Each text type, i.e. ACADEMIC, FICTION, NON-ACADEMIC, NEWS and other written texts, together with SPOKEN ones, has been investigated in relation to the absolute frequencies of the three lemmas.

Lemma	Academic (RF)	Fiction (RF)	Non-academic (RF)	News (RF)	Other written (RF)	Spoken (RF)
disease	4,994 (281.2)	328 (17)	2,591 (95.2)	645 (60.9)	1,831 (72.7)	291 (24.3)
illness	940 (52.9)	310 (16)	1,170 (43)	382 (36)	702 (27.9)	214 (17.9)
sickness	279 (15.7)	157 (8.1)	286 (10.5)	98 (9.25)	288 (11.4)	101 (8.4)

Table 3.
Absolute and relative frequencies across text-types

As can be seen in Table 3 above, *disease* seems to be the most frequently used lemma throughout all text types constituting the BNC. However, while *disease* has a higher frequency in ACADEMIC, NON-ACADEMIC and

miscellaneous written texts, the highest frequency of *illness* is to be found in ACADEMIC texts, followed by NON-ACADEMIC and NEWS text types. *Sickness*, on the other hand, is mainly found in ACADEMIC, miscellaneous and NEWS WRITTEN texts. This seems to suggest that there is a difference in use and that the three lemmas can be regarded as synonyms only in particular text types. The spoken sub-corpus confirms the top presence of *disease*, followed by *illness* and *sickness*.

Table 4 below presents the statistical tests we carried out to measure dispersion and distribution across text categories.

	[DISEASE]	[ILLNESS]	[SICKNESS]
Mean frequency	79.02	30.2	11.66
Standard deviation	81.23	14.22	6.76
Juilland's D test	0.61	0.82	0.78

Table 4
Dispersion across text categories

Based on mean frequency, the most frequent word is again *disease*; however, it is not as evenly distributed across the text categories investigated. Juilliand's D for *disease* is the lowest due to academic text bias.

In the following paragraphs, the results of the collocation analysis of the three lexical items will be discussed.

4.2 Semantic Profiles

4.2.1 Disease

Overall, there were 10,680 occurrences of the lemma *disease*, of which 8,855 were used in singular form and 1,825 in plural form. As aforementioned, the analysis of the lemma and both word forms of *disease* started out with the first 150 collocates identified by the two statistical measures LL and LR. Collocates appearing in both lists were selected for semantic analysis. This reduced list contained 89 collocates in the case of the lemma, 93 in the case of the singular form and 135 collocates in the plural form. These results reveal that although the number of occurrences of the plural form was about a quarter of the number of occurrences of the singular form, the plural form seems to be more productive in terms of the number of different collocates.

The qualitative analysis of the collocational patterns of *disease* in terms of the semantic preference of its different forms yielded interesting results. As can be seen in Table 5, altogether, 23 different semantic categories were identified among the collocates of all forms of *disease*.

SEMTAG	Semantic category	Number of collocates (*disease*)	Number of collocates (*diseases*)	Examples
BODY PART	reference to body part, location of the disease in the body	31	11	*arterial, bladder, gall, gastrointestinal*
NAME	specific name of a disease	17	10	*Alzheimer, Creutzfeldt-Jakob, diabetes, legionnaire, HIV, malaria, measles*
QUALITY	quality of the disease	15	20	*addictive, autoimmune, malignant, severity, acute*
TRANSMISSION	how the disease develops or is transmitted	9	20	*communicable, inherited, transmitted, blood-borne, insect-borne, infectious*
EFFECT	reference to symptom, effect of the disease	8	10	*obstructive, suffering, symptoms, ulcerative, die*
CAUSE	reference to cause of the disease	4	8	*alcoholic, pathogenesis, accidents, cause caused, viral*
WHO/WHAT	reference to whom or what is ill	2	8	*sufferer, elm, animals, cattle, fish, plant*
FUNCT. WD	function words	0	10	*against, among, from, which*
TREATMENT	treatment	2	8	*diagnosis, clinics, treat, treatment*
QUANTITY	quantity	1	8	*sporadic, prevalence, rare*
PREVENTION	prevention	0	5	*combat, drugs, prescribed, prevent*
SYNONYM	sg that is similar to disease	1	3	*pests, illnesses*
DIFFERENT	different	0	3	*different, various, other*
EXAMPLE	example	0	2	*such as, including*
SPECIFICITY	specific	0	2	*specific, certain*
GEOGRAPHY	geography	0	2	*tropical, Western*
LACK	reference to lack of food	1	1	*malnutrition, starvation*
LEGAL	reference to legal documents	0	1	*acts*
TIME	reference to age	0	1	*childhood*
OTHER	one specific case	0	1	*incidence*
RISK	risk	0	1	*risk*
ONSET	start a disease	1	0	*onset*
INSTITUTION	Institution	1	0	*centre*

Table 5
Semantic profile of *disease* based on collocational analysis.

The highest number of categories was identified among collocates of the plural form, which directly corresponds to the highest number of collocates for *disease*. As can be seen from the data, most of the collocates of the plural form are associated with the quality and characteristics of the disease and how it develops or is transmitted. The singular form, however, is most frequently associated with a body part affected by the disease. Collocates of the singular form are often the specific names of diseases, for example,

Alzheimer, Creutzfeldt-Jakob and *malaria*. The semantic categories of symptoms or effects, represented by such collocates as *affecting, crippling, deaths* and to a lesser extent the causes of the disease, illustrated by collocates such as *causes, fungal, parasitic, smoking, viral,* are equally represented among the collocates of the singular and plural forms. Interestingly, a few function words, such as *and, are, as, from, of, these*, collocate with the plural rather than the singular form of *disease*. Fewer collocates are related to treatment, for example, *cure, diagnosis, hospital, treat, treating*, prevalence (quantity) of the disease, for instance, *common, many, multiple, number* and who or what is ill, examples of which include *animals, cattle, fish, horses, human* and *patients*.

4.2.2 Illness

Overall, there are 3,718 occurrences of the lemma *illness,* of which 3,208 are used in the singular form and 510 in the plural form. Based on a comparison of the numbers of collocates generated by the two statistical measures (LL and LR), the first 136 collocates for singular form and 47 collocates for plural form were included in the analysis. The collocates were classified into 11 semantic categories. The relevant semantic categories with a few examples are presented in Table 6, below. The semantic categories include, for example, specific names of illnesses, such as *schizophrenia*, *asthma* or quality, characteristics of an illness, such as *dangerous illness*, *serious illness, common illness*, as well as the cause of the illness, such as *injury* or *HIV*-related illnesses.

The collocation analysis of *illness* has revealed that there are no significant differences between the singular and plural forms. Mostly, *illness* is used in relation to psychological and mental illnesses, for example, *mental illness*, *psychosomatic illness*, *depressive illness*. In addition, the collocates of *illness* also describe the symptoms and effects of such illnesses, for example: *life-threatening illness, depressive illness, long-term illness.* Another important semantic category of the collocates of the semantic profile of *illness* is about dealing with *illness(s),* for example, *treat illness(s), recover from illness(s), prevent illness, overcome illness, cope* with *illness, diagnose illness.*

SEMTAG	**Semantic category**	**Number of collocates (*illness*)**	**Number of collocates (*illnesses*)**	***Examples***
FUNCT. WD	Function word	34	19	*or, often, with*
EFFECT	Reference to symptom and/or effect of the illness	28	5	*life-threatening, flu-like*
DEAL	Dealing with illnesses	22	4	*treat, recover, diagnose*
TYPE	Type of illness in general	16	5	*mental, recurring*
QUALITY	Quality, characteristics of illness	15	4	*dangerous, serious*
CAUSE	Reference to cause	11	4	*injury, HIV-related*
WHO/WHAT	Reference to people	9	3	*patient, childhood, family*
NAME	Specific names of illnesses	5	1	*schizophrenia, asthma*
BODY PART	Reference to body part, location of the disease in the body, function of the body.	4	1	*brain, respiratory*
TRANSMISSION	How illness develops, transmits etc.	4	1	*infectious, enteric*
LACK	Reference to lack of resources (usually food)	1	0	*poverty*
INSTITUTION	Institution	1	0	*hospital*

Table 6
Semantic profile of *illness* based on collocation analysis.

4.2.3 Sickness

Overall, the BNC indicates 1,209 occurrences of the lemma *sickness*, of which 1,205 are used in the singular form and only 14 occurrences in the plural form.

The procedure for the semantic analysis of *sickness* was identical to that adopted for the analyses of *disease* and *illness*. As for the previous lemmas taken into consideration, collocates were detected with both LL and LR; this resulted in a reduced list of 79 collocates, of which 77 can be found in the case of the singular form *sickness*, and 4 in the case of the plural one. All the occurrences have been grouped into 10 semantic categories, as can be seen in Table 7, below.

SEMTAG	Semantic Category	Number of collocates (*sickness*)	Number of collocates (*sicknesses*)	*Examples*
LEGAL	Reference to sickness in legal terms: job and sickness allowance, benefits, rights, insurance, social security etc.	36	0	*absence, absenteeism, allowance, invalidity, rates (of sickness absence), statutory sick pay*
TYPE	Reference to type of sickness	15	1	*altitude, decompression, radiation, spells*
FUNCT. WD	Function words: preposition/conjunction	8	1	*among, and, for, from, of, overall, through*
CONDITION	Human condition	5	1	*age, death, health, ill,* also metaphorical: *the State's sicknesses*
DEGREE	Degree of sickness	5	0	*bout, levels, days, grade*
EFFECT	Reference to symptom, effect	3	0	*diarrhoea, effects, symptoms*
QUALITY	Quality of the disease, characteristic of a disease	3	0	*chronic, acute, long*
SYNONYM	Synonym	1	0	*illness*
SUPERSITION (in literary contexts)	Popular belief, superstition	0	1	*evil*

Table 7
Semantic profile of *sickness* based on collocation analysis.

The data suggest that the collocates of the singular form are associated with a wider range of semantic sets. The most prevalent use is linked to those types of *sickness* which may affect professional life, for example, *allowance*, *benefits*, *insurance*. This seems to indicate that sickness is often used to refer to a state of health in a legal sense. Even when the semantic profile refers to the type and degree of sickness, both its symptoms and characteristics are related to the types of sickness that affect employment life from an insurance or pension-system point of view.

The plural form of *sickness* co-occurred with only four different collocates. These refer to a type of condition that has led to sickness and is related to superstition. It must be said, however, that the plural forms occur in a spoken classroom context in which people are commenting on a literary text. However, given the extremely low frequency of *sickness* in the plural form, far-reaching generalizations cannot be drawn.

Overall, our findings reveal that there are considerable differences between the frequencies, numbers of collocates and which text types the selected near synonymous lexical items are frequently used in. In addition, it was found that the collocational patterns of the examined lexical items also show marked differences in the numbers of collocates and their semantic preferences. This corresponds to earlier studies that suggest that near

synonyms exhibit different collocational behaviours and semantic preferences (Xiao and McEnery 2006). Previous research on the collocational patterns of lexical items suggests that individual word forms of the node word often collocate with different words (Hoey 2005; Gledhill 2000; Sinclair 1991; Tognini-Bonelli 2001). The health-related words examined show similar collocational behaviours, as their singular and plural forms exhibit different collocational behaviours in terms of both collocates and semantic preference.

5. Conclusions

The study has revealed that the three terms under investigation, despite being seen as near synonyms, differ in their collocational behaviours and therefore exhibit different semantic preferences. Overall, *disease* was found to be the most frequent of the three terms. Several semantic categories were identified among its collocates and there is a marked difference in the number of semantic categories associated with the plural and singular forms, the plural form being more productive in terms of both the number of individual collocates and semantic categories. These categories indicate that the plural form shows a semantic preference for how diseases are spread and what they are like. At the same time, the singular form has a semantic preference for the semantic category of body parts and the names of types of disease, as for instance indicated in excerpts (1), (2) and (3) taken from the BNC (emphasis in the original texts):

(1) Not only are you much more likely to die from **lung cancer** or **heart disease**, but other illnesses highlighted in this booklet, including cervical cancer, are associated with smoking (A0J_1708)

(2) Now one of the **auto-immune diseases** that has been recognised is erm unusual baldness — it's called alopecia (KRF_662)

(3) Er, this particular **disease**, **Alzheimer's** disease was identified by Jim, who was the deputy mayor, a member of ours, who spoke to you earlier in the week […] (KM0_688)

The most prevalent use of *illness* is related to psychological/mental illnesses and, in particular, to their symptoms and effects. The collocation analysis of *illness* revealed that words that it co-occurs with are more about dealing with *illness(s),* unlike *disease* and *sickness*, as excerpts (4)-(7) seem to suggest:

(4) Officers had undertaken a review of the policies in both mental **illness** and **mental** handicap in response to the 1975 White Paper and 1976 priority services recommendations. (CS7_1027)

(5) The Royal Commission on Mental **Illness** and **Mental** Deficiency introduced the concept of guardianship, and the Mental Health Act 1959 gave the guardian wide powers of control. (EA1_151)

(6) Caring for a relative with a progressive, relapsing **illness** or **terminal** condition makes future related adverse events almost certain (J14_1281)

(7) It is beyond dispute that advances in medicine and improvements in living conditions have enabled individuals who at previous times would not have survived severe **illness** or **chronic** handicaps to live on, perhaps with some disability, into their seventh, eighth, and ninth decades. (CK_187)

These categories indicate that the singular form of *illness* has a semantic preference related to society, as it indicates how a disease may affect professional life and how this has to be regarded within pension-system or insurance-benefit contexts, as can be seen in example (8), below:

(8) If you are disabled by **illness** or **injury** at the time that you enter the agreement, cover will not begin until you return to full time work. (AYP_2356)

As far as *sickness* is concerned, the semantic categories identified among its collocates show that the term is mainly used in the singular form, which is the only productive one:

(9) AIDS touches areas of **sickness**, death and personal behaviour. (A01_513)

(10) **Sickness**, diarrhoea and some drugs may stop it working, and extra precautions must be used. (A0J_366)

(11) If you are entitled to a **sickness allowance** under the occupational sick pay scheme, SSP is paid as part of that sickness allowance. (HD2_2055)

The plural form is found in the spoken corpus only and is linked to a comment made in relation to a textual analysis:

(12) Erm, but in fact, she's she's missed the third sentence and, where she said that the rose has withstood many **sicknesses** and **evils**, erm, whereas in fact, what it says is it withstands and succours against sicknesses and evils, which is a totally different element.

Strangely enough, the collocates in the BNC seems to tell a different story for *sickness* from those indicated by the dictionaries we consulted: sickness has apparently less to do with a temporary condition and more with a socio-economic one.

Our study seems to bring to light fine-grained differences in the meanings of lexical items, making it possible to achieve a far higher level of precision of a sense of disambiguation, for example, in reference works, such as dictionaries, and better matching in an interlingual quest for equivalence.

The method of semantic profiles, as outlined in this paper, has proved to be especially effective for the task at hand. As we have seen in the excerpts above, *disease* is in most cases accompanied by its scientific or popular name, and is mainly used in relation with the body parts affected by it, also to indicate its characteristics and (particularly in the plural form) how it is transmitted. By contrast, the term *illness* indicates the type of disease, its health effects and how these have to be treated. *Sickness* is preferred when the speaker wants to show the effect the disease has on professional life (allowance, benefits, insurance etc.), and therefore has more social implications.

Although this study has some limitations, primarily the small number of lexical items examined, it nevertheless offers interesting insights into how semantic profiles can be outlined. This can be helpful for research in translation studies and language teaching, as it grants lexical and semantic completeness for the terms under investigation.

As a next step, knowing that the BNC is just *one* of the available resources we have at our disposal and acknowledging that the lingua franca of medicine is English, in both its British and American varieties, we aim to compare the findings resulting from a similar investigation to be carried out on COCA to see the extent to which, if any, (a) meanings can vary and (b) whether meaning variations associated with these items depend on the interactants (i.e. professionals/laymen).

Acknowledgements: We would like to thank the anonymous reviewers for their helpful and constructive comments that greatly contributed to improving the final version of the paper. We are also grateful to the editors for their support, generous help and impressive feedback. We would also like to thank Vaclav Brezina, from Lancaster University, for his great support when we decided to write this paper.

Bionotes:
Stefania M. Maci is Associate Professor of English and pro Vice Chancellor for Education at the University of Bergamo. Her main research focusses on medical and tourism discourse, Critical Discourse Analysis, Corpus Linguistics, Discourse Analysis, Multimodality.

Dr Reka R. Jablonkai works at the University of Bath. She is Director of Studies of the MA TESOL programme. Her research interests include corpus-based discourse analysis, corpora in English language teaching and learning, ESP and EAP, and intercultural communication. She is grant holder of the BAICE Seed fund.

Marek Łukasik, PhD, assistant professor at the Department of English Studies, Institute of Modern Languages, Pomeranian University in Słupsk, Poland. Head of Laboratory for

Modern Methods in Applied Linguistics. His main research interests include metalexicography, terminography, corpus linguistics and specialised translation.

Sophiko Daraselia is Doctoral Researcher and Teaching Assistant at the University of Leeds, UK. Her research interests include corpus linguistics, lexicography and morphology, with particular focus on corpus design and construction and corpus analysis methods, software tools and application of corpus methods in lexicography.

Daniel Knuchel is a PhD Candidate in German Linguistics at the University of Zurich. His main research interests are Discourse Analysis, Corpus Linguistics and Semantics.

Authors' addresses: stefania.maci@unibg.it; reka.jablonkai@gmail.com; marek.lukasik@apsl.edu.pl; mlsd@leeds.ac.uk; daniel.knuchel@ds.uzh.ch

References

Baethge C. 2008, *The Languages of Medicine*, in "Deutsches Ärzteblatt International" 105 [3], pp. 37-40.

Baker P., Gabrielatos C. and McEnery T. 2013, *Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word 'Muslim' in the British Press 1998-2009*, in "Applied Linguistics" 34 [3], pp. 255-278.

Baker P., Gabrielatos C., Khosravinik M., Krzyżanowski M., McEnery T. and Wodak R. 2008, *A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press*, in "Discourse & Society" 19 [3], pp. 273-306.

Barron A. 2003, *Acquisition in interlanguage pragmatics*, Benjamins, Amsterdam.

Bednarek M. 2008, *Semantic preference and semantic prosody re-examined*, in "Corpus Linguistics and Linguistic Theory" 4 [2], pp. 119-139.

Brezina V., McEnery T. and Wattam S. 2015, *Collocations in context: A new perspective on collocation networks*, in "International Journal of Corpus Linguistics" 20 [2], pp. 139-173.

Busse D. 2015, *Sprachverstehen und Textinterpretation*, Springer VS, Wiesbaden.

Candlin C.N. and Candlin S. 2003, *Health care communication: A problematic site for Applied Linguistics research*, in "Annual Review of Applied Linguistics" 23, pp. 134-154.

Collins English Dictionary. https://www.collinsdictionary.com/dictionary/english (17.3.2018).

Crystal D. 2002, *The Cambridge Encyclopedia of the English Language*, Cambridge University Press, Cambridge.

Davies M. 2004, *BYU-BNC*. (Based on the British National Corpus from Oxford University Press). https://corpus.byu.edu/bnc/ (17.3.2018).

Djulbegovic B. and Guyatt G.H. 2017, *Progress in evidence-based medicine: a quarter century on*, in "Lancet" 390, pp. 415-423.

Evert S. and Hardie A. 2014, *Keyword and collocation statistics: Under the hood of CQPwep*. http://ucrel.lancs.ac.uk/talc2014/doc/Workshop-STATS.pdf (17.3.2018).

Firth J. 1957 [1951], *Modes of meaning*, in Firth J., *Papers in Linguistics, 1934-1951*, Oxford University Press, Oxford, pp. 190-215.

Gablasova D., Brezina, V. and McEnery T. 2017, *Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence*, in "Language Learning" 67 [S1], pp. 155-179.

Gledhill C. 2000, *Collocations in science writing*, Gunter Narr, Tübingen.

Gotti M. 2016, *Variations in Medical Discourse for Academic Purposes*, in Ordóñ-López E. and Edo-Marzá N. (eds.), *Medical Discourse in Professional, Academic and Popular Settings*, Multilingual Matters, Bristol-Buffalo-Toronto, pp. 9-30.

Graddol D. 2006, *English Next. Why global English may mean the end of English as a foreign language*, British Council. www.britishcouncil.org/learning-research-english-next.pdf (17.3.2018).

Gries S. 2008. *Dispersions and adjusted frequencies*, in "International Journal of Corpus Linguistics" 13 [4]: 404-438.

Hardie A. 2012, *CQPweb - combining power, flexibility and usability in a corpus analysis tool*, in "International Journal of Corpus Linguistics" 17 [3], pp. 380-409.

Hoey M. 2005, *Lexical priming*, Routledge, New York.
Hunston S. 1995, *A corpus study of some English verbs of attribution*, in "Functions of Language" 2, pp. 133-158.
Hunston S. 2002, *Corpora in applied linguistics*, Oxford University Press, Oxford.
Loiacono A. 2018, *Expressing Health, Disease and Illness in English*, in Loiacono A., *The Medical Alphabet: An English textbook on healthcare in the digital age*, Matarrese Editore, Andria, pp. 382-405.
McEnery T., Xiao R. and Tono Y. 2006, *Corpus-based language studies*, Routledge, New York.
McInnes B.T. 2004, *Extending the Log Likelihood Measure to Improve Collocation Identification*, University of Minnesota. http://www.d.umn.edu/~tpederse/Pubs/bridget-thesis.pdf (17.3.2018).
Merriam-Webster Dictionary. https://www.merriam-webster.com/ (17.3.2018).
Nelson M. 2006, *Semantic association in Business English: A corpus-based analysis*, in "English for Specific Purposes" 25 [2], pp. 217-234.
Oxford English Dictionary. http://www.oed.com/ (17.3.2018).
Partington A. 2004, *'Utterly content in each other's company' Semantic prosody and semantic preference*, in "International Journal of Corpus Linguistics" 9, pp. 131-156.
Sentel T. and Braun K.L. 2012, *Low health literacy, limited English proficiency, and health status in Asians, Latinos, and other racial/ethnic groups in California*, in "Journal of Health Communication" 17 [3], pp. 82-99.
Sinclair J. 1991, *Corpus, concordance, collocation*, Oxford University Press, Oxford.
Sinclair J. 2004, *The lexical item*, in Sinclair J. (ed.), *Trust the Text. Language, Corpus and Discourse*, Routledge, London and NewYork, pp. 131-148.
Stubbs M. 1995, *Collocations and semantic profiles: On the cause of the trouble with quantitative studies*, in "Functions of Language" 2 [1], pp. 1-33.
Stubbs M. 2001, *Words and Phrases: Corpus Studies of Lexical Semantics*, Blackwell Publishers, Oxford.
Talbott J.H. 1962, *The British Medical Dictionary*, in "JAMA" 180 [11], pp. 990.
Taavitsainen I. 2018, *Meaning-making practices in the history of medical English: A sociopragmatic approach*, in "Journal of historical pragmatics" 18 [2], pp. 252-270.
Tognini-Bonelli E. 2001, *Corpus linguistics at work*, John Benjamins Publishing Company, Amsterdam and Philadelphia.
Xiao R. and McEnery T. 2006, *Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective*, in "Applied Linguistics" 27 [1], pp. 103-129.
Wittgenstein L. 2003 [1953], *Philosophische Untersuchungen*, Suhrkamp, Frankfurt aM..
Wulff H.R. 2004, *The language of medicine*, in "Journal of the Royal Society of Medicine" 97 [4], pp. 187-188.

AIN'T THAT SWEET
Reflections on scene indexing and annotation in the House Corpus Project[1]

DAVIDE TAIBI, IVANA MARENZI, QAZI ASIM IJAZ AHMAD

Abstract – This paper outlines the strategies, rationale and potential uses motivating the construction of the *House Corpus*, a one-million-word corpus that can be accessed by authorised users through the *MWSWeb* site (Taibi *et al*. 2015a) at *http://openmws.itd.cnr.it*. Part 1 illustrates the tools and techniques used to index the corpus data – transcriptions of all 177 episodes in the *House M.D.* series (original US version). In particular, it describes the commercially available *Elasticsearch* (*https://www.elastic.co*), used as an indexing, annotational and search tool. Part 2 explains that this is a multimedia corpus allowing viewings of different *types* of scene. The 6000-plus scenes in the corpus have been annotated in terms of their typological features: *Location type* (e.g. patient's hospital room; medical lab etc.); *Event type* (handover; differential diagnosis; precipitating medical event; patient examination etc.) and *Character Group type* (doctor/doctor; doctor/patient; doctor/caregiver; patient/caregiver etc.). The project envisages the development of various retrieval interfaces, initially *Words*, *Scenes* and *Dialogues*. This will make it possible to carry out searches in terms of *types* of scene and their distribution across the corpus without necessarily involving any other form of searching. Part 3 suggests the value of multimedia corpora in encouraging students to advance their critical discourse analysis (CDA) skills. As an example, it shows how the corpus can illustrate the priority of (inter)textual over lexicogrammatical considerations when formulating tag questions in oral discourse. Finally, the *Discussion* section argues that a typology of scenes appears to be an essential prerequisite for the construction of other types of access to the corpus data in subsequent stages of the project.

Keywords: House Corpus; indexing; scene annotation; functionality planning; CDA.

1. Introduction

This paper is a follow-up to the presentation of the preliminary phases of the *House Corpus Project* at *Clavier 17 – International Conference Representing and Redefining Specialised Knowledge*, held at the University of Bari (30

[1] Part 1 of this paper was written by Qazi Asim Ijaz Ahmad, Part 2 by Davide Taibi, and Part 3 by Ivana Marenzi. Davide Taibi and Ivana Marenzi collaborated in the writing of the remaining sections.

Doi Code: 10.1285/i9788883051531p151

November – 2 December 2017), where the research work so far undertaken was presented in summary form. The *House Corpus Project* is concerned with providing a tool for discourse analysis for university teachers and their students, in particular, those attracted by corpus-based explorations of the discourse structures presented in a contemporary US TV drama. As such, the paper explores assumptions about the goals and methods of corpus construction and classroom use of corpora, suggesting the need for greater alignment of corpus linguistics with the needs of university courses that engage with discourse analysis of contemporary English. To this end, the paper is divided into three parts: Part 1: Semantic Indexing of the *House Corpus*; Part 2: Scene management and scene level access; Part 3: Scene level access, scene annotation and discourse analysis.

One feature described at the Congress that needs to be addressed initially in this paper is its break with traditional descriptions of corpora exclusively in terms of words and word counts. Readers who expect the article to expand on the information given in the abstract – 177 episodes, (about) 1 million words – will perhaps be disappointed as the paper, but not necessarily the entire project, is concerned with the structuring of the search mechanism in terms of *scenes* rather than *words*. Compared to the term *word*, *scene* appears to be a neglected and undefined object within corpus studies despite the fact that scenes are central to the production and critical analysis of countless TV dramas. At the time of writing, a search in *Google Scholar* for the search string "*scenes in corpus linguistics"* produces no hits against twenty-three for "*words in corpus linguistics".* Likewise, a specific search for "*word level indexing*" produces 145 hits, while "*scene level indexing*" produces just five. Four of these make no reference to corpus studies while only one, Salway (2007), mentions the search potential of *manual* scene-level indexing but, alas, only for the purposes of dismissing it as a possibility in the specific field of investigation in question, namely audio description:

> In the past archives such as that of the BBC have been for in-house use only, but the advent of the web creates the demand and opportunity to make them available for public access. A minimal requirement is to store production details such as title, director and genre with every programme and film. More useful though is shot- or scene-level indexing whereby keywords are associated with shots and scenes, enabling users to retrieve precise intervals of video data that match their queries, for example 'find me all scenes showing a woman on a horse'. Creating such indexing manually is prohibitively expensive in many cases, and the challenge of the semantic gap limits the scope for machines to generate keywords by analysis of the pixels in the video data. (Salway 2007, pp. 168-169)

While manual annotation may be inappropriate for the specific needs of audio description, we argue below that it can be beneficial in other specialised fields, such as discourse analysis, especially where it allows the functions of a corpus to be modified through supplementary 'tags' introduced by users. In this

project, we aim to show that the possibilities of creating subprojects within the overall *House Corpus Project* depend on functionalities that allow such user-defined tags to be applied systematically. This, we believe, is an innovative approach to corpus studies which potentially assists teachers who wish to explore discourse in English in their university courses, in particular where this involves characterisation of the differences between spoken and written varieties.

In our experience, all too often corpora are exclusively dependent on word-based search mechanisms which become a straitjacket preventing discourse from being investigated *as* discourse. Indeed, our title *Ain't That Sweet* is an iconic representation of this, linked as it is to the detection of the intertextual features of discourse and specifically to the identification of a scene, as detailed in Part III, where Dr. House sings parts of this famous song's lyrics during a discussion of a patient's medical condition. Sensitivity to intertextual references is not something that word-based search and annotation techniques are noted for. Yet such an approach is central in explaining to students how discourse is rooted in shared culture. Exploring such cultural references assists understanding of discourse in English, which is why we suggest that, learning-wise, student engagement with annotation can be beneficial. Scene-level searching, searching, that is, for scenes that share (con)textual characteristics, is thus a first step towards constructing a corpus that facilitates the exploration of culture-related discourse features.

Our efforts to promote the scene to the status of a searchable unit are inevitably the result of teamwork. The paper is accordingly divided into three parts, with each author describing their contribution. Part 1 describes the construction of a corpus that combines scene-based indexing with traditional lemma-based indexing. Part 2 describes the basic design characteristics of an interface that, in addition to search functionalities, also supports manual scene annotation. Part 3 illustrates how all this constitutes a basis for those classroom projects that subscribe to the discourse analysis goals outlined in this paper.

2. Part 1: Semantic Indexing of the House Corpus

Although *Semantic Indexing* is never easy to define as the concept can be interpreted in many ways and is subject to re-interpretation in the wake of constant refinements and improvements in computational technique, for the purposes of the present article, and indeed the *House Corpus Project*, it may be looked upon as the process of mapping a set of metadata onto the transcripts of each episode of the *House M.D.* series. As such, it is a preliminary step in the goal of building a searchable online corpus. In itself, the task of building a set of metadata, while not requiring any understanding of the meaning or

content of the individual episodes, *does* require considerable understanding and management of the characteristics of three distinct textual entities. These are:

(a) the *transcripts* of each TV episode which have been reconstructed from *source texts*;

(b) the *source texts*, *i.e.* the published *html* documents from which the transcripts have been retrieved; these are more extensive textual units as they include other types of text, most prominently various kinds of advertising;

(c) the *target texts* or *records*, *i.e.* the corpus-ready, machine-readable, searchable transcripts of each episode.

However, the transformation of source texts to target texts is not the only problem to be faced. While experienced readers immediately recognize a transcript as a transcript, closer inspection of episode transcripts (as defined above) will highlight individual differences in the use of transcription conventions by transcribers, for example, the way in which, episode titles and airdates are recorded. The work of semantic indexing presupposes the existence of an *episode template, i.e.* a textual standard to which the target text should conform. The process of semantic indexing is thus one of text modification that attempts to emulate and apply the notion of *episode template* systematically. Whether based on experience, or following explicitly stated guidelines, the enactment of this process requires both knowledge of the organization of texts and computational techniques. In the process of semantic indexing, preparing a transcript for such extraction is accomplished in main three steps: *Content cleaning*, *Semantic Annotation* and *Indexing*, each with various sub-steps, the main features of which are described below.

Content cleaning is the process of textual adjustment that we have outlined above. For the *House Corpus*, it involved turning *html* documents with embedded transcripts into corpus-ready transcripts in various steps, some of which are reproduced in Table 1. The process starts with the retrieval of the *source text* (*Point 1* in Table 1), which is achieved using Jsoup API (1), and subsequently proceeds with the cleaning process itself. The information contained within the <Title>tag of the HTML document is not standardized; each URL may store information differently. Table 1 (*Point 2*) shows five examples of different formats within the <Title> tag.

1. **Fetch the content of each URL.** Content is an HTML document.
2. **Extract episode title, season# and episode# by parsing the <Title> tag of HTML document.**
   a <title>House MD - 1.01 Pilot - House Transcripts**</title>**
   b <title>House MD – 4.13 No More Mr. Nice Guy - House Transcripts**</title>**
   c <title>House – S. EE TTTTTTTTTTT - House Transcripts**</title>**
   d <title>MD - S.EE TTTTTTTTTTT - House Transcripts**</title>**
   e <title>S. EE – TTTTTTTTTTT - House Transcripts**</title>**
3. **Extract the main article from the HTML document of each URL**
   At this point, the HTML document contains transcript along with boilerplate text (advertisements, comments, template, navigational elements and other types of unrelated information).
4. **Extract the "Original Airdate" from the main article**
   Like title, the original airdate is also not standardized. The following are some examples of different formats of air dates from different URLs:
   a) Originally aired Apr 4 2006
   b) Originally Aired MMM DD YYYY
   c) Original Air Date on MM DD YYYY
   d) Original Air Date: : MMMM DD YYYY
   e) Original Air Date: MM DD YYYY
5. **Extract author(s) of the episode by standardizing the non-standardized string "Written by" to "Written by:" string**
6. **Remove unnecessary lines from the main article e.g. disclaimer messages**

Table 1
Steps in Content Cleaning.

Technically speaking, we can summarise the process involved as follows. First of all, dashes "–" (HTML code &#8211) in the title are replaced with the minus "–" (HTML code &#45) sign as some URLs contain dashes and some minus signs within the <Title> tag. Afterwards, if the <title> tag contains the strings "MD -" or "House -", the title of the episode is reduced as a substring starting at index of ("-")+2 and ending at index of ("- House")-1. Otherwise, it is reduced as a substring starting at index of 0 and ending at index of ("- House")-1. At this point, the title string from some URLs' content could contain dashes and dots (with spaces). If dashes are found they are removed from the string, whereas if dots with spaces are found they are replaced with dots. The title of the episode is extracted as a substring starting at index 4. The Episode # marker is extracted as a substring starting at index 0 and ending at index 1, while the season # marker is extracted as a substring starting at index 2 and ending at index 4.

The transcript is then extracted from the HTML document (*Point 3*) using Boilerplate API (Kohlschütter *et al*. 2010: 3), which provides algorithms to detect and remove the Boilerplate text/content around the main textual content of a web page. Other forms of standardization are then applied. For example, *Point 4* in Table 1 relates to the standardization of months as MMM (Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec) as compared with spellings, and above all misspellings, of months found in the main articles of URLs which included: *Janu*, *Febu*, *Marh*, *Apri*, *May*, *June*, *July*, *Augu*, *Sept*,

Octber, *Nove*, *Dece*, *January*, *Feburary*, *March*, *April*, *May*, *June*, *July*, *August*, *September*, *Octobor*, *November*, *December*. Likewise the original airdate is standardized by replacing all cases as "Originally Aired:" and extracted as an index of ("Originally Aired:")+2 while the date was changed from the MMDDYYYY format to the DDMMYYYY format.

The next stage in the *Semantic Indexing process* relates to *Semantic Annotation* using *Named Entity Recognition (NER).* The latter is an information extraction task concerned with finding textual mentions of entities belonging to predefined categories, such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages and so on. NER systems take documents either in the form of blocks of plain text or, more directly, as URLs and transform them into annotated text. In fact, a modified version of *DBpedia Spotlight* was used. *DBpedia* (Lehmann *et al*. 2015) is designed, using the techniques associated with the *Semantic Web* (Berners-Lee *et al*. 2001), to extract structured content from the information created as part of the Wikipedia project. The structured information generated from Wikipedia pages is publicly available on the Web. *DBpedia* allows expert users to *semantically* query relationships and properties associated with Wikipedia resources, including links to other related datasets. As a NER, *DBpedia Spotlight* (Mendes *et al*. 2011) associates (i.e. links) Wikipedia resources to plain text.

Two aspects of the use of *DBpedia Spotlight* need to be highlighted. The first relates to *Transcript annotation*. Given the considerable time required for *DBpedia Spotlight* to annotate large documents, each transcript was split into multiple text blocks of about 20000 characters and then sent to *DBpedia Spotlight* for entity annotation. The resources thus obtained were subsequently merged. Once transformed into a *record* consisting of a transcript (or part of it), annotations, author, episode number etc. formatted in JSON format (Crockford 2006), each transcript was ready to be indexed. The second aspect relates to *Scene-wise annotation*, a defining feature in the *House Corpus Project*, which requires the possibility for each scene in an episode transcript to be extracted as a separate entity. The method used is a *Regular Expression* of the form **?((?i)cut .*?)\\]|CUT TO:* which, translated, relates to any characters followed by the string "cut" or "Cut", followed by more characters and a closing square bracket "]", or just the string "CUT TO:"). Scene annotation is much slower than transcript annotation so that for larger corpora (not the case with the current corpus), the *DBpedia spotlight* service would need to be hosted on a local server for shorter delays.

Indexing is the final stage. This is a procedure whereby a Search Engine creates indices for records, thus allowing it to carry out searches more efficiently (*https://en.wikipedia.org/wiki/Search_engine_indexing*). For this, we used *Elasticsearch* (Gormley *et al*. 2015), a popular search engine. Developed in Java, *Elasticsearch* is released as open source under the terms of

the Apache License. Based on *Lucene*, a free, open-source information retrieval software library, it is distributed, which means that indices can be divided into shards (*i.e.* partitions) and each shard can have zero or more duplicates (by default three for backup and other purposes). Thanks to these features, *Elasticsearch* provides near real-time search capabilities using an HTTP web interface which can be accessed by multiple users. After performing entity annotation, the JSON formatted documents were indexed into an *Elasticsearch* server hosted at the CNR Palermo, Italy (*http://openmws.itd.cnr.it*). A final consideration is the fact that indexing is such to allow the exclusion of some parts of the records from the indexing process. Thus, before indexing, it is essential to determine the right mapping for the index (JSON structure where the searchable fields, data types and sub types of fields are declared). For the *House M.D.* series, the default mapping of *Elasticsearch* was used whereby all the fields are set as analysed (*i.e.* searchable). However, *separate* indexes were created for full transcript documents (episodes) and split documents (*i.e.* those based on scenes).

3. Part 2: Scene Management and Annotation

So far, the major focus in *House Corpus Project* has been on encouraging the capacity of university students, many in the very first years of degrees in language studies, to explore the grammar of English in ways that extend beyond the very basic frameworks acquired during years at school. This is achieved by encouraging engagement with the functions of specific lexicogrammatical structures in the scripted discourse of a well-known TV series. As well as supporting *Search functions*, the interface is also designed to allow students to perform further annotation of the corpus under the guidance of teachers. In a project designed to encourage participation in the manual annotation of corpora, the planning of scene-level indexing and of functionalities ideally needs to be built on the premise that the division of the 177 episodes into 6000-plus scenes, carried out in the preliminary stages of the project, opens up the possibility of creating maps of *scene types*. Intuitively, our experience of TV medical drama series suggests the following sequence of events: 1) a person is unexpectedly taken ill and rushed to hospital; 2) the patient is stabilised and the doctors attempt to establish the cause of the illness; 3) complications such as a condition's rarity or concealment of information lead to improper diagnosis; 4) the true cause is eventually uncovered (in this TV series by Dr. House) and the case resolved; 5) the patient, from being on death's doorstep, miraculously recovers and lives happily ever after.

The likelihood that different discourse structures will operate in different parts of an episode will be apparent, even from this basic sketch. For example,

we would expect the present tense verb form *faints* to appear as part of the "stage directions" of an opening scene in which a character falls ill but for the past tense verb form *fainted* to appear in a history-taking and patient examination scene, shortly afterwards, where doctors get to grips with what actually happened to the patient. This pattern does in fact emerge: the form *faints* appears in *Scene 1 of Episode 9, Season 7*, and, as predicted, in a stage direction, while *fainted* appears early on in three episodes (*Season 3, Episode 18 Scene 03; Season 7, Episode 03, Scene 03; Season 8, Episode 14, Scene 07*). However, intuition is not enough to explain why *faints* also occurs in the resolution phase of an episode (*Season 2, Episode 16, Scene 25*) and *fainted* occurs as part of the complication phase (*Season 6, Episode 20, Scene 18*).

While word searches, as the *faint* example show, are a basic premise for the mapping of the various scenes, it is useful to turn matters around and make a scene search the starting point for discovering, for example, the list of verbs typically used in a specific *type* of scene, regarding which it is much harder to make intuitive predictions. Such maps are likely to be useful in supporting the work of various categories of potential users: apart from specialists in media discourse (Baldry 2016), they include all those interested in medical discourse, not just students and teachers of medical English, but also researchers and others developing or participating in specialist classes for medical translation and interpreting (Bianchi 2015). Furthermore, a typology of scenes appears to be an essential prerequisite for the efforts to construct a dialogue level of access, which, in its turn, is likely to be of benefit, for example, to those working in fields such as pragmatics and multimodality. However, in keeping with our primary goal of assisting student annotators in the discovery of discourse patterns within teacher-led projects, the focus has been on providing functionalities that make such manual annotations possible.

Put another way, the interface had to be as intuitive as possible, simplifying the *how-does-it-work* aspects of searching and annotating the corpus, while at the same time encouraging the desire to use the tool as a way to reflect on how the grammar of English is actually used in the production of discourse. To this end, though separate, the interface's *Search* and *Annotation* functionalities are essentially specular, making it easy for students to test out the annotations they make immediately, all part of the process of encouraging discussion of their results with others, a vital aspect of the interface's capacity to stimulate identification of distinctive discourse patterns.

For the purposes of illustrating the interface's characteristics, we will first illustrate the *Search* interface, before describing the corresponding functionalities in the *Annotation* interface. As the first column in Figure 1 shows, the *Search Panel* interface allows selections to be made in terms of individual words or expressions made up of more than one word (*Word Panel*) that can be searched for in terms of the type of scene in which they appear. The

second column in Figure 1 shows the searchable scene characteristics available (*Scene Panel*) relating to the way discourse is shaped and constrained by: a) *Location Type*, e.g. taking place within a hospital setting or elsewhere; b) *Event Type*, e.g. involving patient examination and history-taking, surgery, or, as shown in the example, a case discussion; c) *Interaction type* – currently restricted to scene closures (Coccetta 2019).

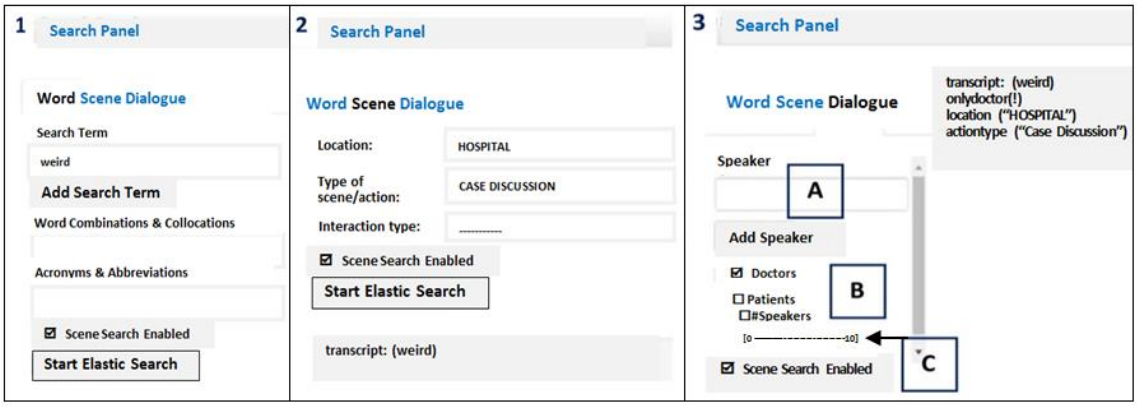


Figure 1
The three-part Search Panel.

The third column in Figure 1 shows a final panel (*Dialogue Panel*) relating to the interactants in the discourse. As the boxed letters show, this allows the user, for example, to select scenes in terms of: (a) specific speakers (*Box A: Speaker*); (b) categories of speakers (*Box B: Speaker Category*); (c) number of speakers in a scene (*Box C: Speaker Number*). As the first column in Figure 2 further illustrates, an entry for CUDDY and HOUSE in the *Speaker* textbox, requires the use of the *Add Speaker* function (*Box A*), plus selecting the *Speakers Box* (*Box B*), setting *0-10 Slider* to *2*, (*Box C*) and finally selecting the *Scene Search Enabled* box (*Box D*). This is all that is needed, apart from clicking the *Start Elastic Search* button (*Box E*), to identify the 169 scenes in which the *only* interactants are House and Cuddy. Cuddy is House's boss and there are many memorable scenes in which they confront each other alone so that, an expert user will want to learn more about the distribution of these scenes across the series. This function is carried out by the *Scene Summary* tool (*Box F*) illustrated in the second column of Figure 2. This generates a table which, although presented here in a clipped form for reasons of space, still identifies fluctuations in scene counts across seasons for this pair of characters and, indeed, shows that this type of scene disppears in the very last part of the series. To understand why, the user can check up on each individual scene using the *Web* tool (*Box H*).

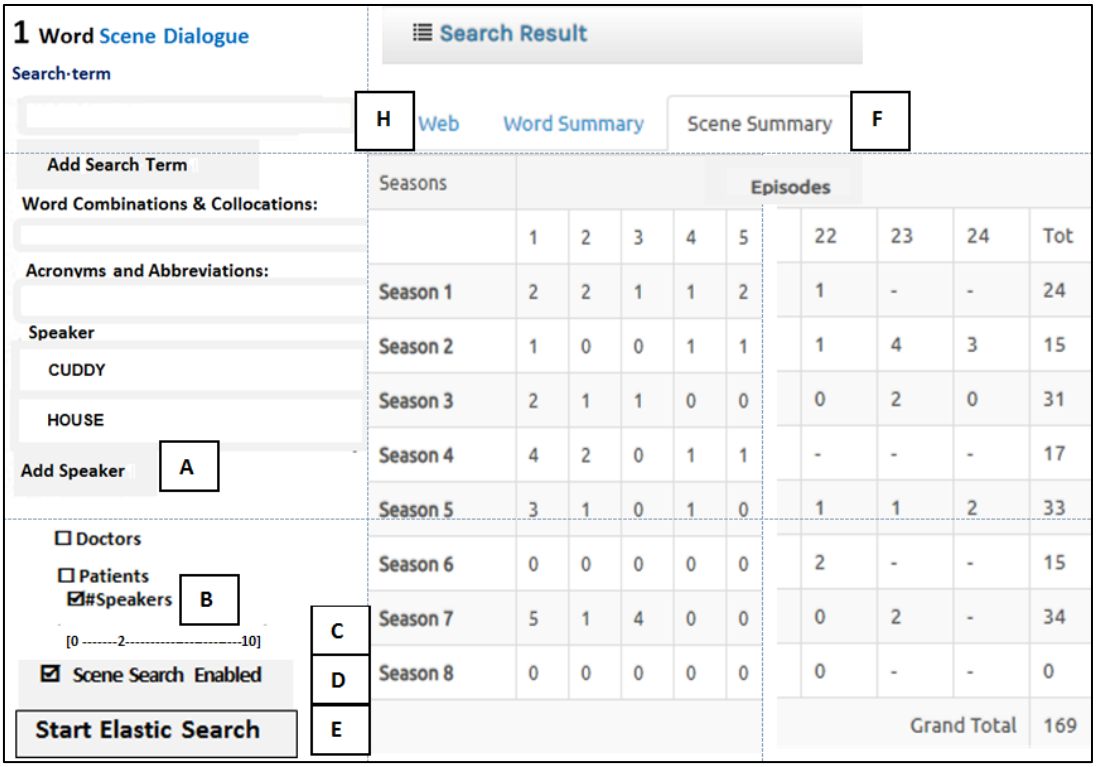


Figure 2
A Scene Search relating to two named interactants.

Additionally, thanks to the work on annotation undertaken by student annotators it is now possible to select scenes in terms of *Character Groups*, for example, scenes, which include doctor-only verbal interactions or scenes characterized by doctors' interactions with patients. As the *Scene Summary* functionality in the *Search Results Panel* in Figure 2 shows, the distribution of such scenes during the unfolding of the TV series varies considerably. The combination of this functionality, and the *Character Groups* functionality, sets up the possibility for teacher-led projects to be carried out that are sociolinguistic in nature and which might well be concerned with speaker distributions and the reasons for such variations in the various episodes and at different points in the overall TV series. While the tools already available are enough to enable such a project to be undertaken, other projects will require adjustments to the interface. For example, Figure 3 illustrates the fact, mentioned above, that each of the scenes identified in Figure 2 can be accessed via the *Web* functionality, the leftmost option in the *Search Result Panel's* main menu and marked as *Box H* in Figure 2. The scenes that Figure 3 reproduces are the first (*Example A*) and the last (*Example B*) of the Cuddy/House face-offs in the series' first season. Both examples in Figure 3 illustrate the constantly conflictual relationship existing between these two characters mentioned above that constitutes a major source of entertainment in the series as in other TV series (Baldry 2016). In this respect, *Speaker initiation* is high

on the *to-do* list as regards functionality development as the search (Figure 2) which detected 169 scenes involving Cuddy/House interactions does not currently distinguish between those initiated by Cuddy and those initiated by House, a distinction that may well reveal differences in the incidence and circumstances of their confrontations.

The search subpanels in *House Corpus Search Panel* interface can be used separately or in combination. For instance, *Case Discussion Scenes*, can be subcategorized into those occurring *within* a specific *Character Group* (e.g. doctors only) and those occurring *between* a specific *Character Group* (e.g. doctors) and a specific individual (e.g. a patient or caregiver) named in the *Speaker Box*. Equally, the *Public/Private* distinction helps clarify why some House-Cuddy confrontations take place before intimidated patients but others occur more privately. The *Word Summary* functionality reports the distribution of searched-for words. As Figure 3 shows, searches need not be lemma-based but in many cases benefit from the inclusion of words. Had the word *job*, which appears in both scenes in Figure 3, been included in the search, the *Word Summary* tool would have shown the distribution across the series of the sixteen scenes with this combination of word and scene features. Additionally, in the individual scenes returned, the target word would have appeared in red as illustrated in many other examples in this article.

SEASON: 1 - Episode: 01 - Pilot - Scene: 04

CUDDY: I was expecting you in my office 20 minutes ago.
HOUSE: Really? Well, that's odd, because I had no intention of being in your office 20 minutes ago.
CUDDY: You think we have nothing to talk about?
HOUSE: No, just that I can't think of anything that I'd be interested in.
CUDDY: I sign your paychecks.
HOUSE: I have tenure. Are you going to grab my cane now, stop me from leaving?
CUDDY: That would be juvenile.
[Both enter the elevator]
CUDDY: I can still fire you if you're not doing your job.
HOUSE: I'm here from 9 to 5.

Example A

SEASON: 1 - Episode: 22 - The Honeymoon - Scene: 29

CUDDY: I want to run something by you.
HOUSE: [loudly] I will not have sex with you! Not again! Miserable, that first time. All that desperate, administrative need –
CUDDY: Stacy's husband is going to need close monitoring at the hospital. And since we can definitely use her back here, I've offered her a job. General Counsel.
HOUSE: Did she say yes?
CUDDY: She said only if it was okay with you. [HOUSE starts to walk off as The Rolling Stones' "You Can't Always Get What You Want" plays ironically in the background.] Yes or no?
HOUSE: Fine. Good.

Example B

Figure 3
Retrieved scenes: the first and last in this TV series where Cuddy and House clash.

As Figure 4 shows, access to specific scenes is made possible using the *Web* functionality, the first option in the *Search Result* Panel. This produces a list of scenes below the heading *Results for Web pages* ranked chronologically in the form of hyperlinks. Mouse selection of the final item in each hyperlink

displays the scene in question. In this example, the words *Scene 03* in Figure 4, when clicked, will display the scene reproduced in *Example A* in the top part of Figure 3. *Dialogue Summary*, the final functionality in the *Search Result* Panel, is designed to quantify the frequency of specific types of exchange patterns but currently has the status of a *yet to be activated* option with characteristics to be defined on the basis of user feedback.

The division of the *Search Interface* into three panels is thus compatible with further customization and addition of new panels meeting the needs of teachers wishing to carry out specific student projects. Some of these have already been incorporated. Hence Figure 1 includes the possibility for searches to be carried out in relation to medical acronyms and abbreviations (Loiacono, Tursi, *this volume*). Equally, provision has been made for *Interaction types* to be included, currently implemented in terms of adjacency pairs (Coccetta 2019). Another project, involving dentistry students, is dealing with the annotation of behavioural verbs such as *cough* and *breathe* and will presumably lead to further adjustments of the interface.

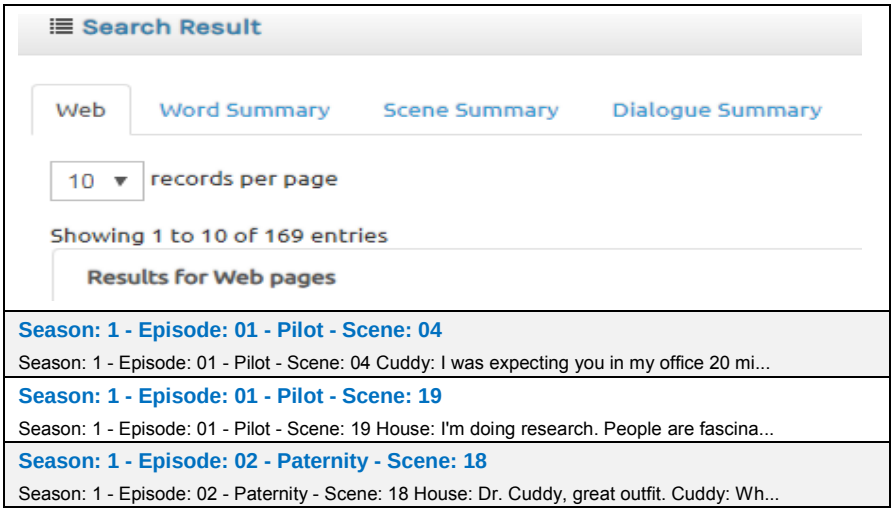


Figure 4
List of scenes relating to Cuddy-House verbal exchanges.

Access to *Search* and *Annotation* functionalities is restricted through the *Profiling system.* The *Manager* functionality illustrated in Figure 5 shows the three steps required to provide groups of students with access to specific functionalities while excluding others. In the example shown, selection of the *Manager* functionality (first column, *Box A*), leads to a *Group Name* functionality (second column, *Box B*, in this case *Student Annotators*) followed by the addition (when so required) of a *Username* and *Password* (third column, *Box C*). Initially, this was a straight choice between *Searching* and *Searching and annotating* (*i.e. Transcript Editor*, third column, *Box D*), but a *Timepointing* functionality described below (see Figure 7a) was subsequently

added. Further customisation, the result of user suggestions and analytics *i.e.* recordings of typical user-corpus interactions, will obviously be undertaken where appropriate.

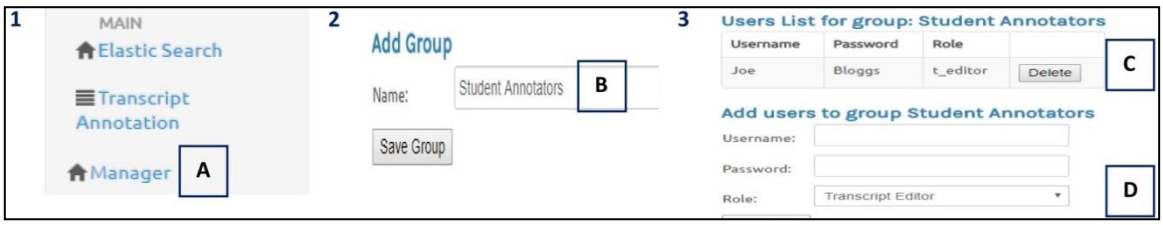


Figure 5
Profiling system.

A partial illustration of the *Annotation Panel's* replication of the *Search Panel* interface is given in Figure 6, which exemplifies the icon-assisted possibilities for annotating specific scenes in relation to intra and extra hospital *Locations*, as well as undecided cases, i.e. those where a decision for annotators is hard to make. Having browsed through the scene in question (shown out of focus in the background), the annotator chooses from a list of over 50 extra-hospital settings used in this series, an easy choice in this case as the scene (*Scene 1, Episode 8, Season 1*) takes place in a classroom. The chosen option remains when the list is closed.

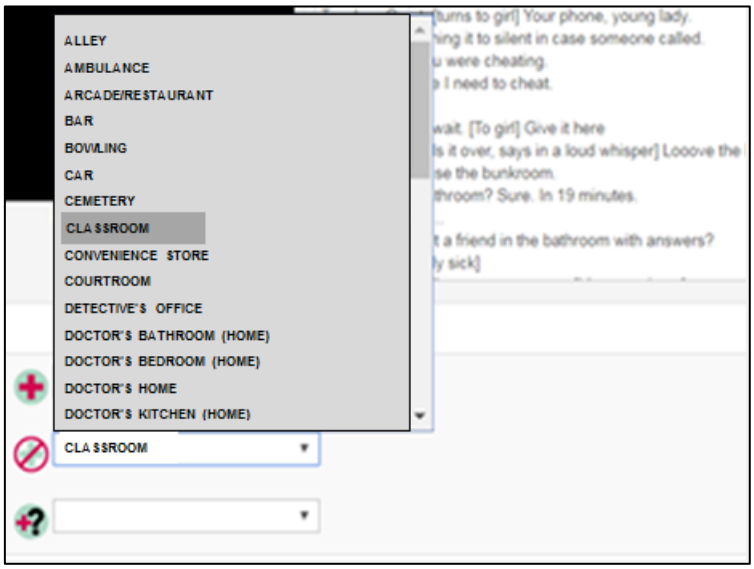


Figure 6
Some options for the annotation of scenes.

The access pathway to individual scenes is through a standard tree structure as illustrated in the various columns in Figure 7a. When the first column in Figure 5 is compared with the top-left hand corner of the first column in Figure 7a, it will be noted that the *Annotation interface* has changed. Thus, in this

configuration, in contrast to providing access to the *Transcript Annotation* functionalities illustrated in Figure 6, access is given to the very different *Timepoint Annotation* functionalities.

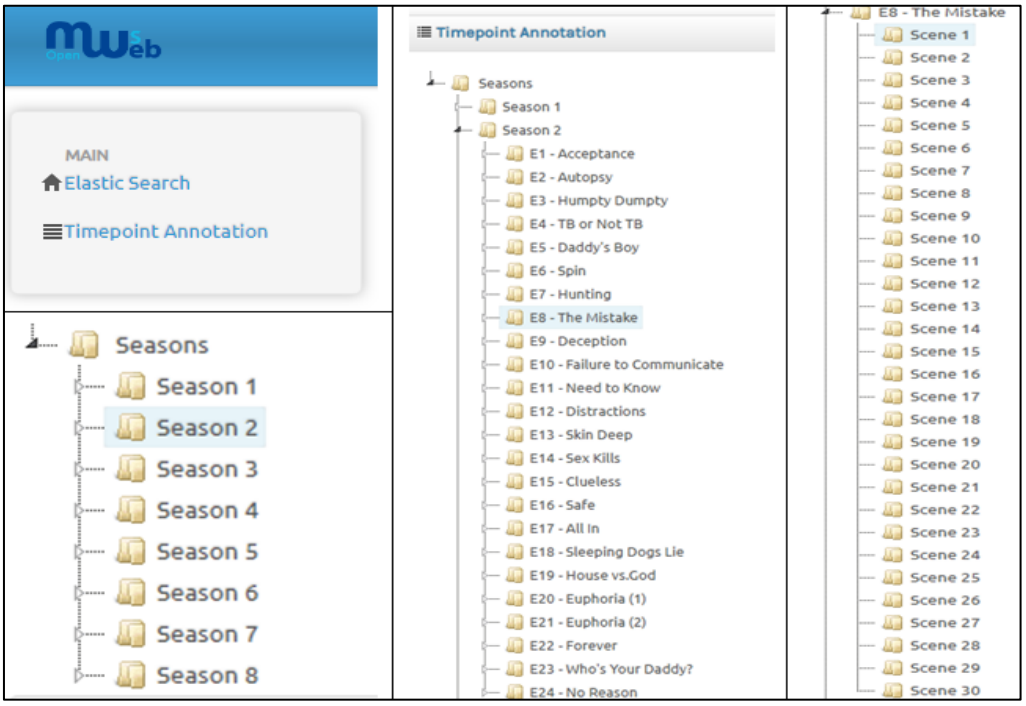


Figure 7a
Accessing annotation options for the link-up between scene reading and viewing.

The latter highlight the role student annotators can play in the work of associating scene transcripts with the corresponding video scene, as illustrated in Figure 7b. As the third column in Figure 7a shows, access to the scene shown in Figure 6 (*Scene 1 Episode 8, Season 2*) has been provided through the same access pathway but, as Figure 7b shows, the *Annotation* functionalities have changed. *Box A* in Figure 7b shows that the *Annotation* interface allows an annotator to indicate the point in the online video where a specific scene starts (in this case, the opening scene in the video), while *Box B* allows the scene's duration to be recorded.

In the initial state of research these annotations were limited to the *Annotation* interface. However, the now completed timepoint annotation work, undertaken entirely by students, was such to provide the data needed to support a corresponding *Search* functionality that allows an end-user to view, as well as read, the scenes that a particular search identifies. This takes the form of side-by-side comparisons of transcript and video versions of the same scene and, as Figure 7b shows, is achieved through links to the *DailyMotion* website (*https://www.dailymotion.com*). These links, which comply with the copyright restrictions stated on this website, encourage deeper investigation into the

relationship between grammatical forms and their discourse functions, thanks to the possibilities of hearing as well as seeing the actors play out their lines.

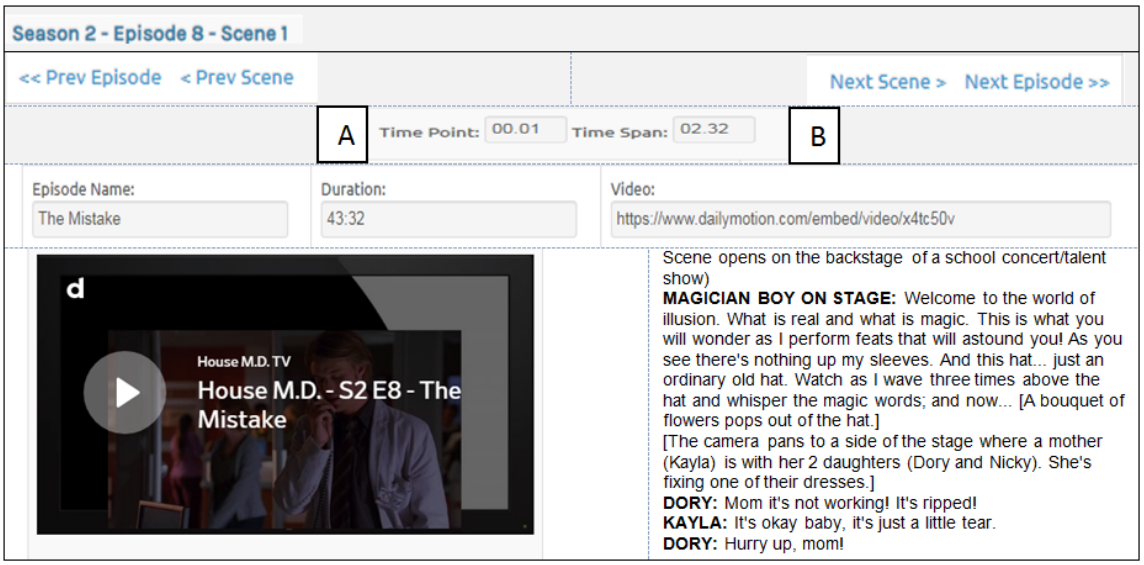


Figure 7b
Using the annotation options for the link-up between scene reading and viewing.

4. Part 3: Scene annotation and discourse analysis

TV drama series such as *House M.D.* offer many opportunities for a better understanding of the functions of scripted TV discourse in English. Given the need for high audience impact, the social and medical contexts chosen by screenwriters adopt a great variety of grammatical forms matched by an equally extensive variety of discourse functions. Take, for example, tag questions. Their use characterises oral discourse interactions in all varieties of English, although with surprising variations, in particular as regards frequency, in different parts of the English-speaking world (Tottie, Hoffmann 2006). While it is quite possible to find side-by-side spoken and written examples of tag questions in close-captioned *YouTube* films, detecting them may be likened to a hunt for microscopic needles in giant haystacks. The *House Corpus* instead finds examples easily and quickly. *Box A*, in the first column in Figure 8, shows how different tag and associated structures can be searched for, using, at the same time, the *Question Tag enabled function* indicated by *Box B*. This function eliminates tag-like forms which are not in clause-final position. As well as illustrating the system affordances, the examples in the second column of Figure 8 – multiple searched-for forms highlighted in red in a specific scene – also show various aspects of tag question patterns that can be used as a model by teachers in their illustration of the grammatical *vs.* discourse properties of tags in oral discourse in English. Thus, the first example shows a positive *anchor* (*it's*) and a negative *tag* (*isn't it*), while the second exemplifies the opposite polarity: a

negative *anchor* (*you're not*) combined with a positive *tag* (*are you?*). The last example illustrates a negative tag (*doesn't it*) whose anchor is not another auxiliary but a lexical verb (*sounds*) with the typical subject ellipsis of spoken discourse. Additionally, the example highlights the discourse strategies involved in the use of tags; these often relate to seeking and providing reassurance.

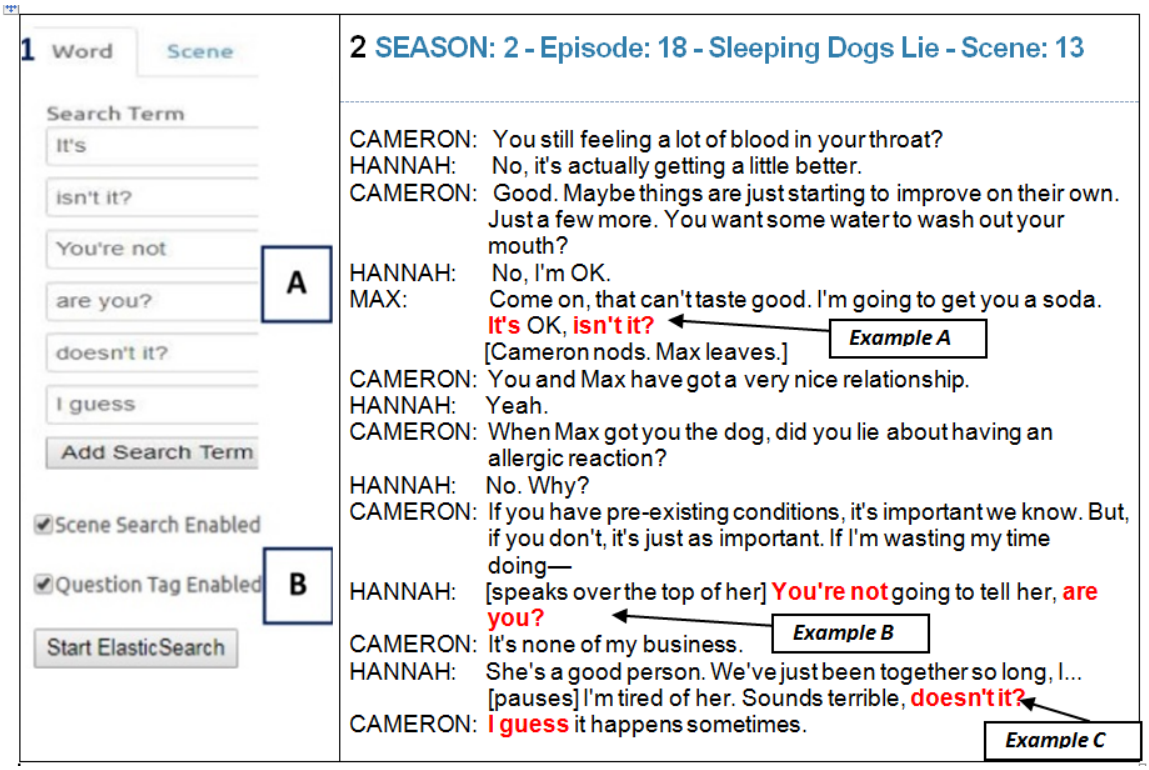


Figure 8
A multiple Question Tag search.

Indeed, most significantly, the added value that scene-based corpus searching brings lies precisely in the characterisation of the different *types* of reassurance sought and provided. In the scene shown in Figure 8, the first example is a request made by Max, the caregiver, for the doctor's agreement. In *Example A*, she seeks and obtains Cameron's reassurance (with a nod of the head) that no harm will be caused if the patient has a soft drink. While this reassurance relates to medical decision-making, the second type of reassurance in this scene regards non-disclosure of information, closely related to the issue which lies at the heart of this episode: professional integrity (*Example B*). Finally, the third example (*Example C*) relates to another type of reassurance concerned with a more personal and psychological plane, in which an experienced doctor allays the feelings of guilt and betrayal that a very sick patient, Hannah, has regarding the desire to leave Max, her companion/caregiver of many years.

Though all this is easily detectable thanks to scene-based searching, manual annotation can, of course, render discourse functions more easily

detectable through specific annotation of question tag functions. However, the examples illustrated in the second column of Figure 8 show that, even without this higher level of annotation, scene-based corpus searches can go beyond typical corpus evidence relating to the frequency of specific lexicogrammatical forms and the ratio of negative to positive tags, as they provide easy access to the discourse functions that specific combinations of forms carry out in specific contexts. Indeed, as well as providing reassurance, tag questions also carry out other functions that demonstrate the need to *hear* and *see* their use in specific scenes in addition to examining them in transcript form. Thus, the pronunciation of a tag such as *"do you?"* – usually glossed as a stressed form when in utterance-final position – will in fact enact differing degrees of markedness according to the speaker's emotional state. Given the nature of drama in general, and House's relationships with his female boss in particular, we can expect that rebuttals rather than reassurance will prevail, as they are part of the conflicts that drive the drama in this TV series along. However, we can never be sure how this will be done. Thus in *Example C* in Figure 9, Cuddy, fishing for a compliment, meets with a rebuttal enacted by the colloquial form *Nope*. In other words, the 'grammatical' expectation, within the turn-taking system of oral and scripted discourse, for tag questions to cue dialogue partners to reply to the question with either a tag-based form of reassurance (e.g. *Yes it is*) or rebuttal (e.g. *No it isn't*) is not always fulfilled. Indeed, none of the take-ups in Figure 9 illustrate the *No, it isn't/ Yes, it is* pattern typically prescribed in rule-based 'grammar' lessons. *Example A* is the closest to such a pattern. It is perhaps easy to accept a response such as *Very* (Figure 9, *Example B*) as a legitimate and elegant breach of such rules, as this provides a strong form of reassurance. Nevertheless, it is the evasiveness of the final two examples that is particularly striking, so much so that, as *Example D* in Figure 9 shows, the original transcriber was so surprised that he or she wrote the bracketed words *[no answer]* immediately after the *isn't it* tag. Indeed, in contrast to the final example, Figure 9, *Example E* – where the *listener* takes evasive action and declines to respond to the tag question – *Example D* in Figure 9, is, instead, an instance of self-directed talk, a case where the current speaker breaks the next-speaker selection rule associated with tag questions by continuing to talk. Indeed, the speaker, shocked by the photo, is seeking self-reassurance, not reassurance from others. Within a manual approach to annotation, the functions of these four types of reassurance – that we may gloss as *medical*, *professional*, *psychological* and *self-referencing* – can be annotated with functional labels and subsequently searched for.

SEASON: 8 - Episode: 02 - Transplant - Scene: 19

WILSON: This is not an exact process. (to Vanessa) Your small airways are collapsing. You're not getting enough oxygen. I'd like to try forcing an oxygen-rich slurry into your lungs. It should open up the airways and buy you some time until the lungs are ready.
VANESSA: Fluid? In my lungs? Sounds like drow...drowning.
WILSON: It is.
VANESSA: Gonna hurt, **isn't it?**
WILSON: Yes, a fair amount.
VANESSA: No. I'm done.

Example A

SEASON: 5 - Episode: 08 - Emancipation - Scene: 12

FOREMAN: How you guys getting along?
CHASE: And you suddenly care why?
FOREMAN: House was asking questions last week.
CAMERON: I assume Foreman needs us, and he's worried that if we're sniping, we might be distracted.
CHASE: That's kind of insulting, **isn't it?**
CAMERON: Very.

Example B

SEASON: 5 - Episode: 14 - The Greater Good - Scene: 36

CUDDY: What the hell is wrong with you?
HOUSE: Yesterday, you hate me. Today, you're practically weeping on my shoulder. I can only assume that what I'm hearing is your aunt flow telling me...
CUDDY: When I was being a jerk, you suddenly act human. But when I act human, you turn back into a jerk.
HOUSE: Guess our cycles aren't matched up yet.
CUDDY: This is your way of saying you accept my apology, **isn't it?**
HOUSE: Nope, this is my way of saying you were doing a crappy job before; you will do a slightly crappier job now.

Example C

SEASON: 5 - Episode: 03 - Adverse Events - Scene: 36

LUCAS: She didn't buy it.
HOUSE: Damn. So you didn't get anything.
LUCAS: Nothin'. We probably overstepped. You're really not the cheerleader type.
HOUSE: On the other hand, I figured she probably wouldn't figure me as the "photoshopping a photo and planting it in an obscure college paper" type either.
LUCAS: Heh. Yeah, about that. I took a little trip to your alma mater.
HOUSE: You took a little trip 150 miles.
LUCAS: Online, by phone. I meant I did research. [House sits and picks up a guitar. They start improvising together.] That's a real photo, **isn't it?** [no answer]. Wow, that is humiliating.

Example D

SEASON: 5 - Episode: 13 - Big Baby - Scene: 13

HOUSE: We got a green light. Go draw the patient's blood.
THIRTEEN: Why?
HOUSE: To see if it clumps in the cold.
THIRTEEN: She's making you confirm your theory before you treat?
HOUSE: She approved the bath. Just thought we ought to do a test to confirm.
KUTNER: That's more of a yellow light, **isn't it?**
TAUB: So she lets you nuke the patient, no problem, but makes you jump through hoops to give her a bath?

Example E

Figure 9
Contextualizations of the *isn't it?* tag question.

It could be argued that an interface specifically designed to look for *anchor* and *tag* sequences would represent an improvement over the current *Tag Question Search* function which merely allows searches for *tag questions* (and not their anchors) to be made. In this respect, a further consideration is that structures exist in English that have the same form and final position in

utterances as tags. However, as the *JENNIFER: Stop it, will you?* example (*Season 7, Episode 20, Scene 22*) shows, such forms have no anchor. They are not a *You won't stop it, will you?* type of structure and do not express reassurance-seeking functions. On the contrary, they are typically demands for something to be done in moments of crisis or conflict and with a degree of insistence bordering on anger. If we add House as speaker into the search using the *Dialogue Panel* in the manner illustrated above in *Part 2*, it immediately becomes clear that four of the five examples of this type in the *House Corpus* are uttered by House and that this structure is associated with his role as team leader in medical emergencies, as Figure 10 illustrates.

SEASON: 2 - Episode: 23 - Who's Your Daddy? - Scene: 32

HOUSE: Pretty much normal. Liver function tests are good.
CRANDALL: Thanks, G-man.
HOUSE: What makes you think you'd be a good father?
CRANDALL: I don't know. Feels right. It feels good.
HOUSE: Well, at least you've got a good reason.
CRANDALL: It feels good is a good enough reason. [Leona begins to choke.] What's happening?
HOUSE: She's choking, she can't breathe. Get him out of here, **will you?** Out! [grabs random instruments] Quick, the curtain! You're breathing on your own, choking's normal. I lied to him, I ran a paternity test. Your lie was a bad one. He is your dad. [to Crandall] We're even.

Figure 10
Contextualization of will you?

However, there is considerable complexity associated with detecting *anchors*, and highlighting them for easy user identification. Tags are constructed from a closed set of grammatical items, listed in Table 2, consisting of: (a) auxiliary and modal verbs with either negative or positive polarity (a distinction marked in Table 2 with a slash) and (b) personal pronouns plus *there* and *one*.

Am/Ain't	Can/can't	Did/Didn't	Is/isn't	Was/wasn't
Are/Ain't	Could/couldn't	Had/hadn't	Must/mustn't	Were/weren't
Is/Ain't	Do/don't	Has/hasn't	Shall/shan't	Will/won't
Are *or* Am/aren't	Does/doesn't	Have/haven't	Should/shouldn't	Would/wouldn't

Table 2
Tag Question Set.

However, their anchors belong to a far less restricted set of grammatical structures (see *Example C* in Figure 8). Indeed the anchors for *do, does* and *did* tags, and their negative counterparts, belong to an open-ended class of lexical items. Moreover, in some cases, no anchor will be present as a result of ellipsis (see *Example A* in Figures 9 and 11). The last line in the first column of Table 2 also includes the tag *am I* as in *I'm not here, am I*. Like the *ain't* form, this breaks with the basic pattern as the 'reverse' form, *I'm here, aren't I*?, requires different morphological selections compared with other cases where the order of negative and positive forms can, in theory, be swapped freely. Whether they

are is another matter: some forms such as *can't* are so frequent that they appear in every episode of *House M.D.*, while the forms *mustn't* and *shan't* appear in none, thus *de facto* reducing the number of potential tag question *type:token* ratios to be tabulated and possibly presented, for example, in classroom teaching.

It will always be possible to find ways of automatically detecting and highlighting the ties between anchors and tags, and thus provide a resource that illustrates significant patterns of cohesion in oral discourse. However, as further suggested below in the *Discussion Section*, within the logic of student engagement with annotation advocated in this paper, it seems more appropriate to carry out manual annotation of anchors that encourages students to explore the 'conflict' between 'grammar' rules and 'discourse' rules and understand that they are two interdependent aspects of the overall process of meaning making.

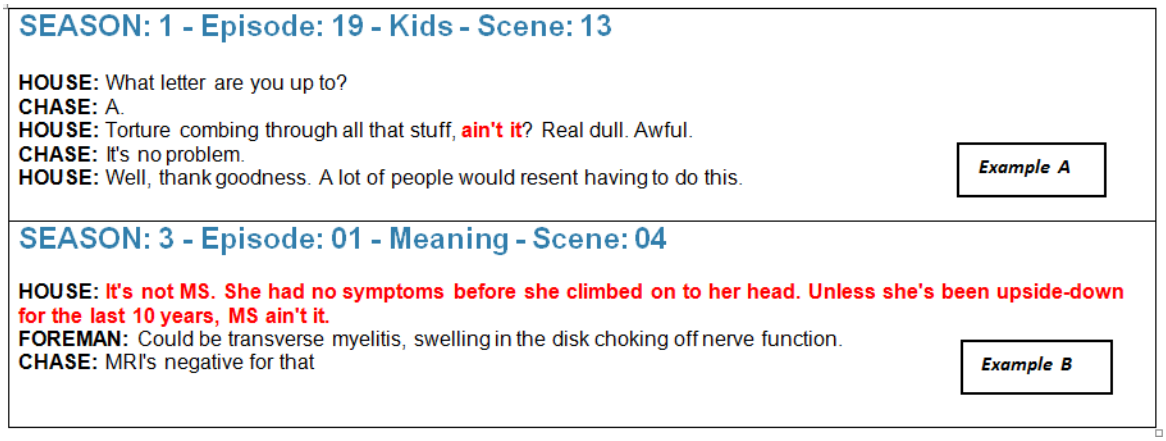


Figure 11
Contextualization of *ain't it.*

Given the limited resources so far available in this project, of more immediate concern have been the investments required to link up transcript scenes with their corresponding video scenes. Even so, it is worthwhile re-affirming the significance of prosodic features in distinguishing tag *look-alikes* from the real thing and hence the fundamental importance of comparative side-by-side readings and viewings that specialised multimedia corpora like the *House Corpus* make available. Alongside forms as such as *isn't it?*, considered 'standard' forms in oral discourse across many varieties of English, there are other forms viewed as substandard whose credentials are rarely presented in English language lessons in schools. As Cheshire (1991) points out, *ain't* is a frequent non-standard form of American and British English, not inflected for person and number, with five 'standard English' equivalents: *haven't, hasn't, (a)m not, aren't* and *isn't*. Figure 11 presents two examples of *ain't it* in the *House Corpus*, the first of which (*Example A*) is a tag question while the second

(*Example B*) is not. Viewings of the two scenes illustrated in Figure 11 show completely different intonation and stress contours that are in keeping with the different functions performed.

Figure 12 shows a scene where *ain't,* eschewed in written discourse in English, is once more used, this time with reference to a jazz era song: *Ain't he sweet.* Like its stablemate, *Ain't she sweet,* it epitomises the freedom of expression and defiance vis-à-vis expected grammatical and discourse strategies that characterise all songs. The song has been sung in many parts of the English-speaking world and recorded by a multitude of singers, including such household names as Nat King Cole, Frank Sinatra and the Beatles, promoting *ain't* as a form characteristic of informal varieties of English. It was thus only to be expected that Milton Ager and Jack Yellen's lyrics (https://lyricsplayground.com/alpha/songs/a/aintshesweet.html) would come to be woven into the *House M.D.* series. Figure 12 reproduces the scene where the devious and deviant Dr. House sings two lines from this song mixing medical lexis with jazz-era colloquialisms, thereby breaking the conventions of case discussions and differential diagnosis – as well as illustrating the need for corpus studies to find ways of detecting intertextual references. Naturally, manual annotation is one such way.

SEASON: 2 - Episode: 09 - Deception - Scene: 22

HOUSE: "See him walking down that street, so I ask you very confidentially, **ain't he sweet?**" Epstein-Barr titers are through the roof, most common viral cause of aplastic anemia. So what I'm saying is, "Just cast an eye in his direction, oh me oh my, **ain't that perfection**?"
FOREMAN: Fetal hemoglobin's also elevated.
HOUSE: Eh, **just a wee bit.** Could indicate –
FOREMAN: Uh, you see that in sickle-cell.
HOUSE: Not all sickle-cell patients are black.
FOREMAN: None of her other blood panels showed any sign of sickle-cell, which means either something's changed drastically since yesterday, or this isn't her blood.
HOUSE: Of course it is! Metaphorically. Look, I couldn't do the tests. I tried, there wasn't enough blood left over. If you just let me do the biopsy...

Figure 12
Contextualization of *ain't he sweet* and *ain't that perfection.*

Songs and singing are essential to any TV drama series. *House M.D.* is no exception. *House M.D.,* like many TV series, is characterised by the constant presence of music and song, in its affirmation of American language and culture (Law 2015). As it grows, the *House Corpus* will assist understanding of how grammatical and interactional selections are underpinned by awareness of, and references to, shared culture, songs being just the tip of this iceberg. Quite apart from the possibilities of detecting scenes that include songs, there is a need to reflect on the *textual* functions of songs, and more generally voice prosodics, within TV dramas, a matter that will be investigated in a subsequent phase of research in the *House Corpus Project.* In the *House M.D.* series, linguistic and cultural aspects are constantly referenced and celebrated as is

further underscored in the scene reproduced in Figure 12 with its use of the expression *a wee bit* – universally associated with Scottish speakers – all evidence of the fact that, if all aspects of discourse are to be represented, corpus studies need to entertain the bigger picture of what is culturally shared in the English-speaking world, a picture for which word-based corpus searches are not noted.

5. Discussion

While the number of words spoken in the *House M.D.* has long been established at just under a million (Law 2015), the number of scenes is never mentioned – despite their centrality in any discussion of a TV series. Many type/token ratio analyses for words (Sinclair 1991; Butler 1997), obtained by dividing the number of different words (types) by the total number of words (tokens), have been produced. The procedure has been extensively critiqued with evaluations of a general nature such as Flowerdew's (2012, pp. 13-16) description of the difficulties of identifying types, as well as more specific assessments of their comparative potential in general *vs.* specialised corpus studies such as the work of McEnery *et al*. (2002) in relation to comparison of the BNC and the 100 Corpus of phone transcripts. A search for studies and critiques of type/token ratios for scenes in which the number of different types of scenes is divided by the total number of scenes in TV dramas will, on the other hand, simply draw a blank. Such ratios are the basis for the scene maps described above, a matter which raises the question as to what applications scene type/token relationships are designed to stimulate. There are many potential answers to this question, some involving purely didactic activities such as identifying scenes containing medical acronyms and thus clearly related to the lexical aspects of specialised L2 learning (Loiacono, Tursi, this volume); others instead might be concerned with research activities with no connection whatsoever to language learning or discourse analysis activities, for example, comparisons across different TV medical dramas of specific scene types such as those portraying medical emergencies which might be useful for TV critics. Obviously, there are strong affinities between language learning and discourse analysis activities. For example, corpus annotation of the type envisaged in the *House Corpus Project* obviously promotes active engagement with oral and written discourse in English in ways that encourage indirect forms of language acquisition (Krashen 1982). Many studies have, of course, suggested the significance of video in improving listening comprehension skills in a variety of teaching (Elk 2014), self-learning (Balcikanli 2010; Richards 2015; Takaesu 2017) and testing contexts (Lesnov 2017; Wagner 2010) as well as other more specialized contexts such as those concerned with

the need for specific teacher training (Park, Cha 2013) or general reflection on the use of video in relation to the acquisition of listening and other comprehension skills (Bianchi 2015; Watkins, Wilkins 2011). Even so, to date, few research projects have contemplated the use of a corpus-based methodology that allows specific oral discourse features to be selected and practised with the advantages of precision and selectivity that corpus-based techniques bring. Some of these (Ackerley, Coccetta 2007, p. 353; Coccetta 2011) include multimedia corpus projects that address the cultural and social issues that we have mentioned above.

However, language learning is not what this project is about. Our concern is instead with defining scenes in ways that make them compatible with encouraging student engagement with CDA (critical discourse analysis) within the framework of corpus linguistics. This is the foundation stone on which the *House Corpus Project* is built and why the authors are concerned with the concept of functionality planning and investments in functionalities that bring about new forms of the empowerment that enhance such engagement.

How has such planning affected *House Corpus* R&D? Within the framework of functionality cost-benefit planning, genre selection was the first factor to be considered. The digital age has brought with it new affordances for the simultaneous side-by-side presentation of more substantial units of written and spoken discourse. For example, *Ted Talks* reinterprets the relationship between spoken and written forms in a way that goes beyond traditional subtitling as it allows users to display videos and their transcripts in the same window thus enabling viewers to watch a video and read its transcript simultaneously. Even so, the *Ted Talks* solution only offers: "monologic talk. The camera moves between long or close shots on the speaker, close shots on the projected slides, and long shots on the listening audience" (Bianchi, Marenzi 2016, p. 27). Given that variety is the spice of life, many users, students and teachers alike, will yearn to go beyond the *Ted Talks* 'talk' genre. Although as with many types of lecture, these talks are highly interactive, they do not illustrate the discourse features associated with interactional exchanges in English that characterise many oral discourse genres of English, exposure to which students enrolled in degree courses dealing with English language studies are in desperate need.

Scene analysis is a second example of functionality planning in which cost-benefit analysis was crucial. Our original division into scenes, as recorded in Part 1 of this paper, is based on references to scene cuts described in online transcripts (see also Law 2015) which thus provided a low-cost entry point for the project. However, defining where a scene starts and where it ends affects the way scenes are defined and quantified. Research promoting automatic scene detection has long recognised the difficulties of detecting scene boundaries (Ewerth, Freisleben 2004). Perceptions of what a scene is differ, a factor, which

for better or for worse, constantly needs to be taken into account and, above all, explored in investigations of discourse in English. This explains our cautious use of the expression '6000-plus scenes' when referring to the partial annotation of scenes subsequently carried out by students in the University of Salento in terms of typological features: *location type* (e.g. patient's hospital room; medical lab); *event type* (e.g. differential diagnosis; precipitating medical event; patient examination) and *Character Group type* (e.g. doctor/doctor; doctor/patient; doctor/caregiver; patient/caregiver etc.). Indeed, the number of scenes has already increased thanks to manual annotation carried out by student annotators who have suggested splitting up scenes into smaller ones on the basis of the systematic application of these typological features. In whatever way a scene is defined, there will always be exceptions. For example, putting forward the idea that a scene is defined in terms of a change in location simply raises the question as to what is meant by a change in location and whether, for example, the frequent scenes in *House M.D.* which include multiple flashbacks are to be defined in terms of the *current* or *predominant* location. As such, from a methodological standpoint, promoting the scene to the status of a searchable but manually taggable unit is a liberating factor. At the very least, it enables students to modify the search results produced by allowing them to introduce *their* annotations about scene characteristics in compliance with the objective of promoting corpora as a way into CDA for undergraduate students.

A third example of functionality planning relates to compatibility with the *short course* and *in-spare-time* solutions. Thus, although corpus construction in general remains within the realm of advanced research, a few studies have described and discussed experiments that involve the participation of students. In one such project:

> participants were given access to specialized corpora of academic writing and speaking, instructed in the tools of the trade (web- and PC-based concordancers) and gradually inducted into the skills needed to best exploit the data and the tools for directed learning as well as self-learning. After the induction period, participants began to compile two additional written corpora: one of their own writing (term papers, dissertation drafts, unedited journal drafts) and one of "expert" writing, culled from electronic versions of published papers in their own field or subfield. Students were thus able to make comparisons between their own writing and those of more established writers in their field (Lee, Swales 2006, p. 56).

Such experiments typically rely on a substantial initial training period and are thus often directed to postgraduate students. This is incompatible with the realities of undergraduate training where CDA and corpus annotation cannot afford to overshadow other objectives. Within the framework of the further annotation of a pre-existing corpus, the *House Corpus Project* pursues a policy of creating micro-projects, that are easily manageable within a *to-be-completed-by-the-end-of-term* timescale, or where appropriate, even shorter periods. The major characteristics of this policy are:

1) Minimum-initial procedural training: learning how the system works requires at most a single live demonstration or a manual consisting of a few pages;
2) Targeting of very specific grammatical and discourse features;
3) Promotion of Teamwork: the model is designed for "group project work" among students in the early stages of their academic career; it enhances confidence through awareness that the annotations made add to the value of the corpus;
4) Customisation: the possibility of adding new annotational features that can subsequently be re-used by different groups for different tasks with minimal need for 're-tooling';
5) Teacher management: the teacher conducting a project has considerable control over the project thanks to *profiling tools* and *data analytics* that allow a teacher to monitor the progress of a group of students as well as each student individually.

The *House Corpus Project* envisages the addition of functionalities on an *as the need arises basis*. Indeed, the project depends on two inter-related aspects of interface management, namely the possibility of increasing the number of functionalities but also the adjustments that can be made to existing ones, which includes delegation of *decision-making* about such adjustments to teachers and/or students.

Clearly, this paper reports on the early stages of this project in which frequency data are not be available. Our curiosity is such, of course, that we, too, are eager to learn the ratio of intra- to extra- hospital scenes and whether scenes that occur in an extra-hospital environment are typically shorter or longer than scenes in an intra-hospital environment just as we would like to know the average length of a scene in this TV medical drama genre. Such knowledge would allow us to identify patterns and provide a basis for explaining why such patterns, and exceptions to them, occur. However, from the standpoint of functionality planning our interest lies elsewhere. In the initial stages of the project, as might be expected, the level of delegation was highly restricted. As the use of the *House Corpus* increases, so the pressure to delegate responsibility for the creation and management of functionalities also increases. Let us review these pressures in terms of functionality planning and what delegation of responsibility entails with some concrete examples.

If we return to the issue of speaker distributions within a university CDA short course project with a sociolinguistic orientation, we may note that it is already clearly possible, with the tools already existing, to carry out searches relating to the distribution of scenes *per episode*, *per season* and *per series* in the following ways:

1. *Numerically*: *i.e.* scenes with no speakers, a single speaker, two speakers and so on;
2. *Per individual: i.e.* scenes with specific characters named either in the metadata (e.g. speaker cues) or referenced in the discourse;
3. *Combinations* of these two parameters.

Note, however, that although the corpus is indexed in terms of individually named speakers, the current interface does not fully allow scenes to be identified in terms of speaker characteristics other than speaker name. Minor interface adjustments building on the student annotation functionalities already provided will make it possible to explore the power relationships implicit in interaction in terms of:

1. *Gender: e.g*. a project designed to annotate and explore the ratio of male-only scenes to female-only scenes;
2. *Professional and social standing*: e.g. a project looking into the construction of a *Category group* such as *caregivers* and the interactional expectations and realities associated with this category.

Thus, in the next stage of development, the intention is to create functionalities that allow a greater degree of delegation for a) teachers with respect to the system designers and b) students with respect to teachers in the construction of search categories. Thus, with a view to enabling *Gender* and *Cross category group* annotations, it is intended to:

1. provide the *Dialogue interface* with a *Speaker Group* function that allows new groupings of speaker names to be constructed;
2. allow a teacher to decide whether or not to make the *Speaker Group* function available to students in a project;
3. request students to determine the members of the *Speaker Group* in accordance with a specific project's objectives.

A similar pattern of delegation will likewise allow new annotational subcategories to be added to the pre-existing *Location type* and *Event type* parameters. While such changes require some rewriting of the interface rules, they are well within the bounds of possibility. On the contrary, a similar arrangement, creating a *Word Group* functionality, whereby users define and search for sets of related lexical items within the *Word* interface, would be a time-consuming IT task involving complex search rules and is thus currently not an option being taken into consideration. The issue of tag questions is, indeed, instructive as regards the cost-benefit ratio of investing in certain functionalities and not others in terms of the degree of delegation that can be achieved. Tottie and Hoffman (2006, p. 296) state that, when searching for "entire tags consisting of auxiliary, pronoun, and optional *n't*, we found a total of 200 different combinations, most of them occurring in very low proportions". Thus, as Part 3 has shown, from the standpoint of investment in

learning experiences, delegating the solution to student annotators has many merits. The *Question Tag enabled* functionality currently only available in the *Search Interface* could be added to the *Annotation Interface*, based on a pre-established table of options, such as the one shown in Table 2. This would then allow *manual* annotation of *anchors* to be performed on a scene-by-scene basis by students as an end-of-term class project, with items from the list in Table 2 assigned to different groups. Of course, this raises the issue of the benefits that such a project would bring to the students in terms of exercising their CDA and corpus search & annotation skills, a matter that would have to be decided by the teachers overseeing such a project.

In the current stage of research, it is not entirely possible to predict which functionalities will be required, nor the benefits that the student engagement approach will bring as more data is required, in particular, as regards the value that has been added by associating scenes extracted from the corpus with the corresponding video scenes. The expectation, however, is that the answer to issues of functionality planning lies with data analytics as the recordings of user searches and annotations will provide a better guide to management aspects relating to the delegation, addition and modification of existing functionalities and the cost-effectiveness and benefits to students of further investment in new functionalities.

6. Conclusion

As the article reports, though indexed in ways described in Part 1, the *House Corpus* leaves open the possibility for annotations of a manual nature to be made to specific scenes in the TV series. Through a system of restrictive passwords and other controls, the interface is designed to allow university teachers to carry out specific annotation projects with selected groups of University students in which the scripted discourse of an entire TV series is explored with a view to adding annotations that enrich the value of the overall corpus. As such, while encouraging learning that relates to specific aspects of discourse in English, as illustrated in Part 3 with regard to the use of tag questions, the research reported, in keeping with the training and educational goals promoted by the institutions to which the authors are affiliated, is concerned with the development of online tools that exercise students' ability to acquire critical skills in the description of the discourse of written and spoken varieties of English through a hands-on approach to annotation. From the results so far obtained, promoting students' CDA skills through greater awareness of the characteristics and functions of corpora appears to be a viable proposition.

The project thus raises a basic question about the role of specialised corpora. Are they an end-product to be construed on a par with a printed dictionary for the purposes of consultation or are they to be seen instead as part of a collaborative learning experience in which the corpus itself is subject to the process of modification? From the exposition given above it is clear that the *House Corpus Project* is attempting to provide a strong stimulus in support of the view that specialised corpora can and should drive learning processes through student engagement with annotation and searching. Indeed the tag question example shows that the affordances created by hybrid forms of manual and automated annotation give a new twist to the term *blending learning*. From a procedural standpoint, the tag is identified and highlighted on the basis of abstract search rules enacted by a search engine, while on the contrary, the anchor and the subsequent take-up by a cued interactant could well be part of a student annotation project concerned with investigation of discourse patterns that cause unexpected disruptions to grammatical patterns.

Thanks to the active participation that the annotation of scenes entails, discourse analysis, which might otherwise be considered a rather dull activity, can be turned into a highly active and interactive process of discovery and reflection on descriptive models. Hopefully, the *House Corpus Project* will lead to corpus annotation projects suggested by students themselves. If so, we suspect they may well be directed towards a better understanding of the cultural models hidden in a TV series such as *House M.D.*, most obviously comparisons of expectations about medical services in different parts of the English-speaking world as reflected in answers to questions like *Did the patient lie?* and *Did another doctor screw up?* constantly foregrounded in the *House M.D.* series. Whatever happens in the future, there is considerable satisfaction in knowing that, so far, teacher and student responses to the project have been more than positive.

A final thought relates to the research efforts being made to overcome the risk of corpus studies having little bearing on classroom activities owing to a disproportionate focus on word counts and frequency-based statistics. Our title, *Ain't that sweet,* is a song-like slogan encouraging investments in multimedia corpora that serve the interests of scholars and students by stimulating engagement with the complexities of English discourse. Hopefully, this slogan will work in the same way for others as it has for us.

Acknowledgements: Special thanks to Francesca Bianchi for her assistance and advice with the *House Corpus Project*, in particular as regards engaging undergraduate students in pilot online and offline annotational activities.

Bionotes:
Davide Taibi is a researcher at the Institute for Educational Technology, Italian National Research Council and part-time lecturer at the University of Palermo. His research activities are mainly focused on: Mobile Learning, Semantic Web and Linked Data for education, Open Educational Resources and Learning Analytics.

Ivana Marenzi PhD is senior researcher at the L3S Research Center, Leibniz University of Hannover. Her main research area is Technology Enhanced Learning in support of collaborative and lifelong learning with a special focus on ties between technology and communication. She has published Multiliteracies and e-learning2.0, Peter Lang, Frankfurt, 2014.

Qazi Asim Ijaz Ahmad is a software engineer at the German National Library of Science and Technology (TIB), Hannover. His main work includes development of an open source research information software VIVO, text mining and citation and information extraction.

Authors' addresses: davide.taibi@itd.cnr.it; marenzi@l3s.de; asimijaz@live.com

References

Ackerley K. and Coccetta F. 2007, *Enriching language learning through a multimedia corpus*, in "ReCALL" 19 [3], pp. 351-370.

Balcikanli C. 2010, *Long live, YouTube: L2 stories about YouTube in language learning*, in Shafaei A., Nejati M. (eds.), *Proceedings of the 2009 International Online Language Conference (IOLC 2009)*", Universal-Publishers, Boca Raton, Florida, pp. 91-96.

Baldry A.P. 2016, *Multisemiotic Transcriptions as Film Referencing Systems*, in Taylor C. (ed), *A Text of Many Colours – translating The West Wing*, Intralinea special issue. www.intralinea.org/specials/article/2195 (10.03.2018).

Berners-Lee T., Hendler J. and Lassila O. 2001, *The semantic web*, in "Scientific American" 284 [5], pp. 34-43.

Bianchi F. 2015, *Integrazione e Apprendimento: I prodotti cinetelevisivi come strumento didattico linguistico e culturale per il mediatore e il migrante*, in "Lingue e Linguaggi" 16, pp. 237-263.

Bianchi F. and Marenzi I. 2016, *Investigating student choices in performing higher-order comprehension tasks using TED talks in LearnWeb*, in "Lingue e Linguaggi" 19, pp. 23-40.

Butler C.S. 1997, *Repeated word combinations in spoken and written text: some implications for functional grammar*, in Butler C.S., Connolly J.H., Gatward R.A. and Vismans R.M. (eds.), *A Fund of Ideas: Recent Developments In Functional Grammar*, IFOTT, University of Amsterdam, Amsterdam, pp. 60-77.

Cheshire J. 1991, *Variation in the use of ain't in an urban British English dialect*, in Trudgill P. and Chambers J. K. (eds.) *Dialects of English. Studies in Grammatical Variation*, Longman, London and New York, pp. 54-73.

Coccetta F. 2011, *Multimodal functional-notional*, in Frankenberg-Garcia A., Flowerdew L. and Aston G. (eds.), *New Trends in Corpora and Language Learning*, Continuum, London, pp. 121-138

Coccetta F. 2019, *Old Wine in new bottles. The case of the adjacency-pair framework revisited*, in "Lingue e Linguaggi" 29, pp. 407-424.

Crockford D. 2006, *RFC4627: The application/json media type for javascript object notation notation (json).* [Online]. http://tools.ietf.org/html/rfc4627 (22.10.2019).

Elk C.K. 2014, *Beyond mere listening comprehension: Using Ted Talks and metacognitive activities to encourage awareness of errors*, in "International Journal of Innovation in English Language Teaching and Research" 3 [2], pp. 215-246.

Ewerth R. and Freisleben B. 2004, *Video cut detection without thresholds*, in *Proceedings of 11th Workshop on Signals, Systems and Image Processing*, PTETiS, Poznan, Poland, pp. 227-230.

Flowerdew L. 2012, *Corpora and Language Education*, Palgrave Macmillan, Basingstoke.

Gormley C. and Tong Z. 2015, *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine*, O'Reilly, Beijing.

Kohlschütter C., Fankhauser P. and Nejdl W. 2010, *Boilerplate detection using shallow text features*, in *Proceedings of the third ACM international conference on Web search and data mining (WSDM '10)*, ACM, New York, NY, USA, pp. 441-450.

Krashen S.D. 1982, *Principles and practice in second language acquisition*, Pergamon, Oxford.

Law L. 2015, *House M.D. Corpus Analysis: A Linguistic Intervention of Contemporary American English*, in Li, L., Mckeown, J. and Liu, L. (eds.), *Proceedings of AsiaLex 2015*

Hong Kong: Words, Dictionaries and Corpora: Innovations in reference science, The Hong Kong Polytechnic University, Hong Kong, pp. 230-249.

Lee D. and Swales J. 2006, *A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora*, in "English for specific purposes" 25 [1], pp. 56-75.

Lehmann J., Isele R., Jakob M., Jentzsch A., Kontokostas D., Mendes P.N. and Bizer C. 2015, *DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia*, in "Semantic Web" 6 [2], pp. 167-195.

Lesnov R.O. 2017, *Using videos in ESL listening achievement tests: Effects on difficulty*, in "Eurasian Journal of Applied Linguistics" 3 [1], pp. 67-91.

Loiacono A. and Tursi F., *this volume*.

McEnery T., Baker P. and Cheepen C. 2002, *Lexis, indirectness and politeness in operator calls*, in "Language and Computers" 36, pp. 53-70.

Mendes P.N., Jakob M., García-Silva A. and Bizer C. 2011, *DBpedia spotlight: shedding light on the web of documents*, in *Proceedings of the 7th international conference on semantic systems*. https://dl.acm.org/citation.cfm?id=2063519&dl=ACM&coll=DL (22.10.2019).

Park S.M. and Cha K. 2013, *Pre-service teachers' perspectives on a blended listening course using Ted Talks*, in "Multimedia-Assisted Language Learning" 16 [2], pp. 93-116.

Richards J.C. 2015, *The changing face of language learning: Learning beyond the classroom*, in "RELC Journal" 46 [1], pp. 5-22.

Salway A. 2007, *A corpus-based analysis of audio description*, in Orero, P. and Remael, A. (eds.) *Media for all: Subtitling for the deaf, audio description and sign language*, Rodopi, Amsterdam, pp. 151-174.

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

Takaesu A. 2017, *Ted Talks as an Extensive Listening Resource for EAP Students5*, in Kimura K. and Middlecamp J. (eds.), *Asian-Focused ELT Research and Practice: Voices from the Far Edge*, IDP Education Cambodia, Phnom Penh, pp.108-126.

Tottie G. and Hoffmann S. 2006, *Tag questions in British and American English*, in "Journal of English Linguistics" 34 [4], pp. 283-311.

Wagner E. 2010, *The effect of the use of video texts on ESL listening test-taker performance*, in "Language Testing*"* 27 [4], pp. 493-513.

Watkins J. and Wilkins M. 2011, *Using YouTube in the EFL classroom*, in "Language Education in Asia" 2 [1], pp. 113-119.

Website references

Further information on the tools mentioned in this article

Boilerpipe	https://code.google.com/archive/p/boilerpipe/ (12.11.2017).
DailyMotion	https://www.dailymotion.com (05.04.2018).
Dbpedia Spotlight	https://github.com/dbpedia-spotlight/dbpedia-spotlight (10.09.2016).
Dbpedia	http://wiki.dbpedia.org/ (10.09.2016).
Elasticsearch	https://www.elastic.co (10.09.2016).
Jsoup Java	https://jsoup.org/apidocs/ (10.09.2016).
Lucene API	https://lucene.apache.org/ (10.09.2016).

http://siba.unisalento.it