Les institutions et les médias
De l'analyse du discours à la traduction

Le istituzioni e i media
Dall'analisi del discorso alla traduzione

Édité par
Marie-Christine Jullion, Louis-Marie Clouet et Ilaria Cennamo

Videoimpaginazione: Paola Mignanego
Stampa: Logo

Table des matières

Towards a Linguistic Definition of 'Simplified Medical English': Applying Textometric Analysis to Cochrane Medical Abstracts and Their Plain Language Versions

Christopher Gledhill - Hanna Martikainen
Alexandra Mestivier - Maria Zimina

ABSTRACT

The Cochrane organisation publishes large-scale meta-analyses of clinical trials (Systematic Reviews). These are summarised in two distinct types of text: Scientific Abstracts (ABS), which target experts and researchers, and Plain Language Summaries (PLS), aimed at the general public. Although the Cochrane Style Manual and the PLEACS standards (McIlwain *et al.* 2013) provide some general advice on writing in "simple and accessible English", Cochrane does not define what is meant by Plain English in formal linguistic terms. In this paper, we set out to identify some of the main lexico-grammatical characteristics of the simplified language used intuitively by Cochrane authors. In particular, we use textometric analysis to examine the differences between a representative parallel corpus of ABS and PLS. On the basis of these observations, we provide some tentative recommendations for any future style-guide for writing and editing Cochrane PLS.

Keywords: corpus analysis; lexico-grammatical features; medical abstracts; plain language; textometrics.

1. INTRODUCTION

In this paper, we attempt to define some of the principal features of plain or simplified medical English on the basis of a comparative analysis of two parallel monolingual corpora: a corpus of expert-to-expert Scientific Abstracts (ABS) produced by the Cochrane organisation (introduced below), and a corpus of summaries representing simplified versions of these texts, known as Plain Language Summaries (PLS). Our primary goal is descriptive: to identify the main lexico-grammatical features of PLS. Our hypothesis is that any regularities of expression that can be observed in PLS correspond to some of the underlying features of linguistic simplification, which – we suggest – emerge intuitively when experts attempt to reformulate complex scientific information for non-experts.

Our secondary goal is prescriptive. The Cochrane organisation communicates specialised medical knowledge for both professionals and the general public. Although Cochrane has editorial guidelines (Higgins and Green 2011; McIlwain *et al.* 2013), the organisation has recently begun to re-examine the effectiveness of its communication strategy, most notably by looking at the visual format and the language it uses in PLS, ABS and other communications. One particular issue for the Cochrane editors is how to define 'plain language'. We would suggest that, currently, the Cochrane guidelines say very little about actual linguistic features. The guidelines do mention some general features, such as "use the active voice": however, such advice can be found in many plain language guides. In this paper we set out a series of recommendations which may form the basis for any future set of guidelines that Cochrane wishes to adopt. However, we propose that such recommendations should not be based on intuitions (or what other proposals for plain language should be like), but rather our recommendations should be based on the analysis of a representative corpus of published texts (that is to say, based on those texts that Cochrane considers to be good examples of plain or clear writing).

In the first part of this study, we briefly explore the context of the Cochrane Systematic Review and its accompanying summaries (ABS, PLS), as well as the relevance of such concepts as 'plain' or 'controlled language' to Cochrane's editorial policy. In the main body of this paper, we use textometric analysis and corpus observation in order to identify some of the core linguistic differences between ABS and PLS. The aim here is to establish the specific lexico-grammatical features that are statistically more likely to occur in PLS in comparison to ABS, and can

thus be assumed to be characteristic features of simplified English in this type of text. We focus in particular on the distribution of lexico-grammatical patterns (repeated sequences of Part of Speech n-grams), which we claim correspond to routine patterns of meaning in highly conventional texts such as Systematic Reviews (SR). Finally, on the basis of these observations, we make some tentative suggestions for any future linguistic guidelines that Cochrane may wish to use as part of its style guide for authors of PLS.

2. Cochrane Systematic Reviews: a unique text type

Cochrane is an independent, international network of medical practitioners and other volunteers who produce the Cochrane Database of Systematic Reviews [1]. Cochrane is a leading organisation in the development of high-quality Systematic Reviews in the field of medicine (Grimshaw 2004). The SR is a very particular genre of text: each one presents a meta-analysis of many different studies on a single medical question (e.g. mothers who smoke, different approaches to treat back pain, patients injured after falling out of their hospital beds, etc.). The purpose of SRs is to help clinicians to keep up with the research literature, as well as to provide evidence for policy-makers to judge risks and benefits of health care interventions. According to Cochrane's website, the aim of each SR is to "collate, analyse and critically appraise" data from a selection of studies dealing with a particular medical problem. SRs thus take a critical stance on the quality of the studies they analyse, and are therefore good examples of "highly valued texts" (Halliday and Matthiessen 2004, 4), that is to say texts for which there are high expectations in terms of normative quality and potential social impact.

The unique features of SRs set them apart from other scientific genres such as research articles in peer-reviewed journals. To a certain extent SRs can be seen as an institutional response to the seemingly unmanageable volume of contemporary research. One of the most original features of Cochrane SRs is that each includes both a Scientific Abstract written for experts, and a Plain Language Summary, which is in effect an executive summary of the whole review targeted at patients and caregivers. It is also worth adding that while access to

---

[1] Information about Cochrane as well as its database of Systematic Reviews can be found at: http://www.cochrane.org/ [22/07/2019].

SRs is restricted to paying customers, both the ABS and PLS versions are freely accessible online. This increases the communicative value of ABS and PLS, which can be seen as gateway texts to the rest of the SR. A final particularity is that SRs and their associated ABS and PLS are updated regularly, unlike traditional research articles. All of these features, but most especially the unique relationship between expert ABS and non-expert PLS, make SRs a particularly interesting topic for linguistic investigation.

3. SIMPLIFICATION: SOMEWHERE BETWEEN PLAIN LANGUAGE AND CONTROLLED LANGUAGE

When discussing text simplification, it is traditional to make a distinction between plain language (PL) and controlled language (CL). PL can be defined as "a deliberate linguistic style, adopted with the intent of making expert knowledge clearer and more accessible to non-experts". Thus plain language generally corresponds to a rhetorical or stylistic intention, rather than a specific set of guidelines (Cutts 2013; Pinker 2014). A CL on the other hand can be defined as "a planned language with a named set of lexical, grammatical or pronunciation rules, designed either a) to ease communication in multilingual or technical contexts in which safety and comprehensibility are primary factors, or b) to harmonise the presentation and translation of expert texts for non-expert users". The paradigm examples of CLs are Airspeak (Air Traffic Control phraseology) and Simplified Technical English (STE, Farrington 1996).

There have been proposals for CLs which specifically cater for non-expert medical communication, such as Patient Information Leaflets (Renahy *et al.* 2015). Thus it could be argued that the guidelines for Cochrane PLS might benefit from a set of recommendations that resemble a CL. However, Systematic Reviews and their accompanying ABS/PLS are probably too 'open-ended' to allow for the use of such a restrictive system, except perhaps for a small number of very circumscribed phrases (such as 'final recommendations'). Therefore in this paper, we propose an intermediate approach. On the one hand, the recommendations we set out below are inevitably more systematic (since they are based on descriptive corpus analysis) than the informal guidelines for PL mentioned in the Cochrane style guide. On the other hand, it is clear that our recommendations in no way correspond to a fully-fledged CL.

Before turning to our analysis of PLS, it is necessary to briefly discuss the relative effectiveness of simplified language. There have been many studies in the field of readability metrics which compare original texts and simplified texts using semi-automated measures, as well as user trials comparing reading speeds, recall, etc. (cf. Chervak, Drury, and Ouellette 1996; Stewart 1998; O'Brien 2003; Fonseca 2006). There is some evidence to suggest that, while simplification might result in a lower reading complexity score, paradoxically it may also introduce more complexity, in the form of lexical and grammatical expansions (features which we encounter in the data analysis, below). As Crossley *et al.* (2007) put it:

> [...] simplified texts, in their attempt to elaborate meaning by using simple syntactic constructions, may in fact create sentences that have more constituents and, thus, place a heavier processing burden on the reader than do authentic texts. (Crossley *et al.* 2007, 26-27)

When discussing such issues, Halliday and Matthiessen (2014) make an interesting distinction between "complexity" (the elaborate style of written discourse, involving embedding, subordination, etc.) and "intricacy" (the expansive style of oral discourse, involving parataxis, disruption and so on). However, notwithstanding such issues, in this paper, we remain agnostic about the relative benefits of using simplified language for language processing. Our primary aim here is to describe what happens when the authors of PLS consciously attempt to adopt a 'clear' writing style. As discussed below, such an approach assumes that it is possible to establish best practice on the basis of texts that have already been published, and thus may represent models for emulation.

4. Data, tools and methods

As mentioned above, our hypothesis is that by comparing two sub-corpora (expert-oriented ABS and non-expert-oriented PLS) we can identify some of the typical strategies of simplification adopted by the authors of PLS. For the purposes of the present study, we used a sample of 4,540 SR summaries (which all contain one ABS and one PLS), for a total of 3.3 million words (ABS: 2.1 million words, PLS: 1.2 million words). The texts were written and updated (when new scientific proof became available) between 2005 and 2013.

To compare both sub-corpora, we use the tools and methodology of textometrics. In a previous study, we established a research protocol associating qualitative and quantitative methods for what we call the

"lexico-grammar approach" to analysing specialised discourse (Gledhill, Patin, et Zimina 2017). This framework compares regular patterns across two or more comparable sub-corpora. The first step of the process involves Part-of-Speech (POS) tagging (such as N for nouns, V for verbs, etc.) for each sub-corpus (Schmid 1994). Then, n-gram counts are used to reveal not just the quantities and distributions of POS but more specifically sequences of POS tags known as POS n-grams (such as <Det N>, <Det Adj N>, <V N>, etc.). Some elements in the POS n-gram may be highly predictable ("pivotal") features within longer, more productive lexico-grammatical patterns. Such sequences correspond to chunks of text that express a unitary (albeit abstract) meaning. Our working hypothesis is that in domain-specific texts and specialised genres, such patterns will be (1) easier to spot, and (2) longer and more predictable than in the general language, the reason for this being that in specialised discourse language users tend to use similar communicative strategies, and these strategies are realised by highly regular stretches of language (i.e. long stretches of predictable, but also productive formulaic language). Finally, as a further step in our textometric analysis, we perform the computation of *specificity indices* (showing "over-use" and "under-use" of different POS elements and sequences in different parts of the corpus, cf. Fleury and Zimina 2014).

One interesting feature of contemporary textometric analysis is that, in contrast to previous approaches, textometric analysis provides significant information not only about the raw frequency of linguistic items (or features), but takes information about the relative qualities of the text as a whole into account when identifying the relative importance of these items (or features). In particular, as can be seen in *Figures 1* and *2*, the process of *characteristic elements computation* (Lebart, Salem, and Berry 1998) allows the analyst to detect the significant over- and under-use of any unit under investigation (token, POS n-grams, etc.) in a given part of the corpus, both in relation to its distribution in all parts of the corpus and in relation to the length of the whole corpus. For this computation, a *hypergeometric model* is used (Lafon 1980): this makes use of established statistical tests (Lebart, Salem, and Berry 1991) that can detect, within each part of a corpus, which elements are used frequently as well as those which tend to be rarely used.

Furthermore, different variables (corpus sub-parts) can be analysed within this approach using textometric software. For example, the chart in *Figure 1* shows the main characteristic elements (POS) in the ABS and the PLS sub-corpora. The bars above the horizontal axis indicate over-use of a specific category, and those below the horizontal axis,

under-use; the height of the bar is in inverse proportion to the probability of this over- or under-use (the higher the bar, the more characteristic the element). For example, one can see that proper nouns (labeled NP, NPS), cardinal numbers (CD) and coordinating conjunctions (CC) turn out to be under-represented in the PLS.

Figure 1. – POS elements under-represented in the PLS.

These quantitative analyses were conducted using *Le Trameur* (Fleury and Zimina 2014). This allows for the intersection and visualisation of multiple text annotation layers in various forms of textometric analysis (including multivariate statistics).

Table 1 illustrates some of the findings of our experiments with the ABS and the PLS sub-corpora. From all the repeated POS sequences of the whole corpus, n-grams with the highest global frequency (F > 1,000) and the highest *specificity indices* are displayed (in the table, the symbol ** indicates specificity indices over 50). For each element represented by a POS n-gram (column 1), both the ABS and the PLS sub-frequencies (columns 2 and 4) and characteristic elements diagnostics (columns 3 and 5) are displayed. For each element, the diagnostic consists of a sign and an exponent (Lebart, Salem, and Berry 1998), where:

• +E*xx* corresponds to a *positive characteristic element* to which is associated a probability of the order 1/10*xx*;
• -E*xx* corresponds to a *negative characteristic element* to which is associated a probability of the order 1/10*xx*.

As can be seen, the subset comprises POS sequences of more than three elements (see *Tab. 1*). Interestingly, the longest sequence we can find (IN DT NN IN DT NN) corresponds to the POS sequence <*Prep + Det + N (singular or mass) + Prep + Det + N (singular or mass)*>, which is therefore shown to be significantly salient in the PLS.

Table 1. – Characteristics elements:
highly frequent POS n-grams over-represented in the PLS.

POS N-GRAM	FREQ ABS	SPECIF ABS	FREQ PLS	SPECIF PLS
NN IN DT NN	12,063	-**	8,916	+**
NN IN DT NNS	2,597	-**	2,136	+**
IN DT NN IN DT	1,804	-**	1,665	+**
IN DT NN CC	1,649	-**	1,535	+**
MD VB VVN IN	1,389	-**	1,358	+**
VBZ DT JJ NN	1,315	-**	1,220	+**
NN DT NN IN	1,053	-**	1,058	+**
IN DT NN IN DT NN	1,001	-**	1,001	+**

In order to provide a linguistic interpretation for the statistical data provided by *Le Trameur*, the PLS sub-corpus was indexed using IMS Corpus WorkBench (Evert and Hardie 2011). Using the complex corpus query language (cql) provided by this tool, it is then possible to search the corpus for sequences of word-forms, POS patterns and lemmas or combinations thereof. By concordancing the corpus for the POS n-grams which are over-represented in the PLS (such as the n-gram <IN DT NN IN DT NN>, see *Fig. 2*) and then sorting the contexts, it is possible to zoom in on a specific set of constructions. It is then of course still necessary for the analyst to examine these contexts, but as can be seen in *Figure 2*, although each example is quite different, each one can be characterised as a complex prepositional phrase (embedded in a complex nominal), and thus turns out to be a regular element in a longer – and significant – pattern in the PLS (as discussed in the Results section).

The procedure described above applies only to one sub-corpus: it is of course then necessary to compare sentences in the PLS with sentences or paragraphs conveying the same information in the corresponding ABS. A typical pair of equivalent sentences is set out below (see *Ext. 1*). It demonstrates a preference in PLS for complex nominals with post-modifying prepositional phrases instead of pre-modifying elements (in this case: ABS *permanent posterior teeth* vs. PLS *permanent teeth in the back of the mouth*).
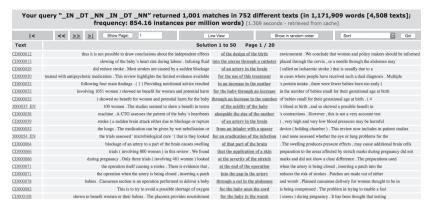
*abdominal aneurysm **in the area of the aorta***. In such cases, this expansion appears to have been adopted with the aim of increasing clarity in the PLS.

Table 2. – Characteristic prepositional n-grams (N ≥ 3)
over-represented in the PLS (F > 1,000).

CHARACTERISTIC POS N-GRAM	FREQ ABS	SPECIF ABS	FREQ PLS	SPECIF PLS
NN_{noun} IN_{prep} DT_{det}	23,723	–**	15,968	+**
NN_{noun} IN_{prep} DT_{det} NN_{noun}	12,063	–**	8,916	+**
NN_{noun} IN_{prep} DT_{det} NNS_{noun_plur}	2,597	–**	2,136	+**
IN_{prep} Dt_{det} NN_{noun} IN_{prep} DT_{det}	1,804	–**	1,665	+**
IN_{prep} DT_{det} NN_{noun} CC_{coord_conj}	1,649	–**	1,535	+**
NN_{noun} DT_{det} NN_{noun} IN_{prep}	1,053	–**	1,058	+**
IN_{prep} DT_{det} NN_{noun} IN_{prep} DT_{det} NN_{noun}	1,001	–**	1,001	+**

5.2. *Over-use of adjectival and adverbial groups*

Adjectives also appear in a number of salient POS n-grams in PLS (including comparative adjectives and adverbs, often post-modified by prepositional phrases, see *Tab. 3*). In the broader context, this tendency often coincides with patterns of expression which involve direct evaluation using an attributive clause, as in *Determining which interventions **effectively** and **safely** prevent patient injuries from their beds would **be beneficial*** or in evaluative statements of the type: *this group of patients may have been **too small** to detect a **statistically significant** benefit or harm in the analysis.*

Table 3. – Characteristic POS/POS n-grams (N ≥ 3)
with over-represented adjectives and adverbs in the PLS (F > 2,000).

CHARACTERISTIC POS/POS N-GRAM	FREQ ABS	SPECIF ABS	FREQ PLS	SPECIF PLS
JJ_{adj}	247,901	–**	131,104	+**
RB_{adv}	62,161	–**	40,185	+**
JJR_{adj_comp}	9,138	–**	7,334	+**
JJ_{adj} IN_{prep} DT_{det}	3,355	–**	2,521	+**
JJ_{adj} IN_{prep} NN_{noun}	3,204	–**	2,476	+**

5.3. *Over-use of the verb be (attributive clauses)*

Attributive clauses followed by a pre-modified nominal (i.e. statements of the type *X **is** a/an ADJ + NOUN*) are also typically found in the PLS sub-corpus (see *Tab. 4*). As with the adjectival patterns seen in 5.2, this often coincides with general purpose definitions, direct evaluations and comparisons, as in *Chronic pelvic pain in women **is a common problem***.

Table 4. – POS n-grams (N ≥ 3) reflecting characteristic presence
of definitions/evaluative statements in the PLS (F > 1,000).

Characteristic POS n-gram	Freq ABS	Specif ABS	Freq PLS	Specif PLS
$VBZ_{verb\_be\_pres\_3rd\_p\_sing}$ DT_{det} JJ_{adj}	1,896	-**	1,728	+**
$VBZ_{verb\_be\_pres\_3rd\_p\_sing}$ DT_{det} NN_{noun}	1,797	-**	1,787	+**
$VBZ_{verb\_be\_pres\_3rd\_p\_sing}$ DT_{det} JJ_{adj} NN_{noun}	1,315	-**	1,220	+**

5.4. *Over-use of verbal groups with infinitives, passives and progressives*

Complex verbal groups (involving auxiliaries and sequences of the type V to V) are also typical in PLS (see *Tab. 5*).

Table 5. – Characteristic verbal n-grams (N ≥ 3) in the PLS (F > 1,000)
reflecting characteristic infinitives, progressive and passive forms.

Characteristic verbal n-gram	Freq ABS	Specif ABS	Freq PLS	Specif PLS
$VVN_{past\_participle}$ TO_{inf} $VV_{base\ form}$	2,998	-**	2,358	+**
NN_{noun} TO_{inf} $VV_{base\ form}$	2,673	-**	2,549	+**
NN_{noun} $VBZ_{verb\_be\_pres\_3rd\_p\_sing}$ $VVN_{past\_participle}$	1,467	-**	2,053	+**
$VBZ_{verb\_be\_pres\_3rd\_p\_sing}$ $VVN_{past\_participle}$ IN_{prep}	1,440	-**	1,588	+**
NNS_{noun_plur} $VBP_{verb\_be\_pres\_non-3rd\_p.}$ $VVN_{past\_participle}$	1,470	-**	1,430	+**

Although the sequences included in *Table 5* involve different types of structures, taken as a whole they suggest a tendency for PLS to use serial verbs involving the infinitive, participles and the passive. The passive is typically used in contexts where a trial or study is presented as part of the Systematic Review, as in *Two randomised studies **are included in** this review*, or the results of trials are summarised (*The review of trials found that more research **is needed to determine** whether heparin in IV fluids is advantageous for neonates*). A further set of structures are involved in embedded infini-

tival and progressive clauses, which often express causation. Examples of both can be seen in: *Bed rails are the most common intervention* **designed to prevent** *patients falling out of bed.*

5.5. *Over-use of verbal groups involving a modal and passive*

Verbal groups involving deontic and epistemic modals also appear to be highly characteristic of PLS (see *Tab. 6*). The most typical of these involve findings and recommendations expressed by *can* and *should*, as in *Antibiotic treatment* **can reduce** *the risk of kidney infections in pregnant women*, as well as passive forms (**can** *be used*, **can** *be treated*, **should** *be used*, **should** *be interpreted*, etc.). An extended pattern involves an additional preposition, which typically expresses 'means' or (abstract) 'location', as in *These women* **can be treated with** *a surgical procedure, prostacyclin and related drugs* **should be used in** *acute stroke.*

Table 6. – *POS n-grams (N ≥ 2)*
reflecting characteristic modals in the PLS (F > 1,000).

Characteristic POS and POS n-gram with modals	Freq ABS	Specif ABS	Freq PLS	Specif PLS
$MD_{modal}\ VB_{be\_base\_form}$	4,469	–**	4,531	+**
$MD_{modal}\ VV_{verb\_base\_form}$	3,935	–**	6,253	+**
$NN_{noun}\ MD_{modal}$	3,403	–**	3,750	+**
$MD_{modal}\ VB_{be\_base\_form}\ VVN_{past\_participle}$	1,896	–**	1,728	+**
$NN_{noun}\ MD_{modal}\ VB_{be\_base\_form}$	1,530	–**	1,354	+**
$MD_{modal}\ VB_{be\_base\_form}\ VVN_{past\_participle}\ IN_{prep}$	1,389	–**	1,358	+**

5.6. *Over-use of personal pronouns, possessive determiners, restricted (defining) clauses*

PLS tend also to characteristically over-use personal pronouns and possessive determiners in a variety of grammatical contexts (see *Tab. 7*). The high use of determiners is associated with animate referents, as in *any adverse effects of antibiotic treatment for the mother or* **her baby**, while pronoun usage coincides with relative pronouns in restricted (defining) clauses such as *patients* **who** *were witnessed to have fallen out of bed*, or *a fall may be defined as an unintended impact with the ground*, **which** *may or may not result in physical or psychological harm.*

Table 7. – POS and POS n-grams (N ≥ 2) reflecting characteristic presence of personal and possessive determiners and restricted (defining) clauses in the PLS (F > 1,000).

Characteristic POS/POS n-gram with pronouns, determiners and wh-words	Freq ABS	Specif ABS	Freq PLS	Specif PLS
$WD_{wh\text{-determiner}}$	5,473	-**	5,627	+**
$WRB_{wh\text{-adverb}}$	4,569	-**	4,058	+**
$WP_{wh\text{-pronoun}}$	2,962	-**	4,353	+**
$PP\$_{possessive\_pronoun} NN_{noun}$	2,041	-**	2,549	+**
$IN_{prep} PP\$_{possessive\_pronoun}$	1,495	-**	1,865	+**
$PP_{personal\_pronoun} VBZ_{verb\_be\_pres\_3rd\_p\_sing}$	1,240	-**	1819	+**
$IN_{prep} PP_{personal\_pronoun}$	1,082	-**	1,407	+**

5.7. *Under-use of cardinals, abbreviations, proper names*

Finally, it is interesting to observe features that are *uncharacteristic* of PLS. As can be seen in *Table 8*, several elements are found to be more typical in ABS than PLS, including proper nouns, cardinal numbers and abbreviations, as in the following example from an ABS: ***Haines 2010** found no significant reduction in the rate of falls in the bedroom… (rate ratio **0.69, 95% CI 0.35**…).*

Table 8. – POS/POS n-grams (N ≥ 2) under-represented in the PLS (F > 1,000).

POS/POS n-gram	Freq ABS	Specif ABS	Freq PLS	Specif PLS
$NP_{proper\_noun\_sing}$	230,104	+**	46,554	-**
$CDc_{cardinal\_number}$	228,038	+**	29,871	-**
$NP_{proper\_noun\_sing} NP_{proper\_noun\_sing}$	40,657	+**	3,263	-**
$NP_{proper\_noun\_sing} CD_{cardinal\_number}$	38,053	+**	1,082	-**
$CD_{cardinal\_number} NNS_{noun\_plur}$	33,396	+**	12,422	-**

5.8. *Keyword analysis*

Space prevents us from presenting all of our results here. However, it is worth noting several elements in the keywords list that are typical in PLS as opposed to ABS appear to be typical of oral discourse. This includes the pronoun *you* and the possessive *your* (see *Ext. 2*), which partly coincide with the results in section 5.6 above (over-use of pro-

nouns) as well as *what* and *how* used in rhetorical questions (see *Tab. 9*). This observation confirms the fact that, as in other types of 'plain language', PLS authors typically use full clauses and the interrogative as a common strategy for their section titles.

Extract 2. – Typical use of you and your in PLSs.

Abstract	Plain Language Summary
Osteoporosis is a condition resulting in an increased risk of skeletal fractures due to a reduction in the density of bone tissue.	Bone is a living, growing part of **your** body. Throughout **your** lifetime, new bone cells grow and old bone cells break down to make room for the new, stronger bone. When **you** have osteoporosis, the old bone breaks down faster than the new bone can replace it.

Table 9. – Use of how/what in PLS titles.

Abstract	Plain Language Summary
Main results	What did the studies show?
Author's conclusions	What is the bottom line?
Main results	How well does it work?

6. From lexico-grammatical patterns to discourse functions

The results we have set out above only identify some of the most typical linguistic features that are typical of PLS as opposed to ABS. However, it is also necessary to establish a link between lexico-grammatical regularities and discourse functions. In particular; the typical lexico-grammatical patterns we discussed in section 5 can be associated with a variety of writing strategies whose overall effect is to render more explicit or to refocus the specialised information expressed in the ABS (and the original SR). As set out below, these observations correspond to two general strategies: (1) explicitation of information-rich discourse (to use a concept from translation theory, e.g. Blum-Kulka 1986) and (2) focus on the human user.

6.1. *Explicitation: unpacking compact nominals*

As mentioned above (section 5.1), we find that compact, heavily pre-modified nominal groups in ABS tend to be 'unpacked' as post-modified nominals in PLS. In functional terms, this corresponds to the breaking down of compact technical terms and their reformulation into more 'user-friendly' formulations with relative clauses and definitions. The following examples (see *Tab. 10*) show how pre-modified nominals in the ABS are reformulated by post-modifying phrases in the PLS.

Table 10. – *Unpacking compact nominals using post-modifying prepositions.*

Abstract	Plain Language Summary
injection site reactions	skin reactions **at the site of the injection**
abdominal aortic aneurysm (AAA)	an abdominal aneurysm **in the area of the aorta**
near the bladder neck and proximal urethra	**around the junction between** the bladder and the urethra
the external urethral sphincter muscle (EUS)	the muscle **at the exit of the bladder** that allows people to control their bladder emptying

It is worth adding here that many post-modifying reformulations in the PLS correspond to a concrete localisation in the human body (with the first noun in the sequence involving an abstract 'facet' of the following noun, such as *area*, *junction*, *site*, *exit*, etc.). Some examples also feature the topicalisation of *people* or *patients* as active participants in these processes (as discussed below).

A related strategy for decomposing complex nominal structures in the PLS involves relative clauses. This can be inferred from the sequences discussed in section 5.6 above. These patterns have a similar structure to post-modifying prepositions, although their function is more specifically to define technical terms, often diseases, or to reformulate research methods and results in terms of processes. *Table 11* gives some examples of this.

A further related structure in PLS involves *with* as post-modifier (e.g. *with a mix of conditions*). Finally, as we discuss below, the use of such post-modifying structures also tends to coincide with more frequent reference to active human participants, as can be seen in the many uses of the relative pronoun *who* (i.e. *people **who** eat a diet rich in antioxidant vitamins*, *smokers **who** try to quit*, *patients **who** received oral immunotherapy*, *babies **who** were large for gestational age*, etc.).

105

Table 11. – Unpacking compact nominals using post-modifying clauses.

Abstract	Plain Language Summary
patients recovering from a stroke, patients recovering from surgery, elderly patients with a mix of conditions	patients **who** had a stroke, patients **who** had surgery, and elderly patients **who** had different types of conditions
a chronic skin disease characterised by inflamed spots	a chronic skin disease, **which** causes spots to occur,
number needed to treat (NNT)	the number of people **that** need to be treated
higher dietary levels of antioxidant vitamins	people **who** eat a diet rich in antioxidant vitamins

6.2. *Explicitation: using simple attributive clauses as definitions*

Our textometric evidence (section 5.3, see *Tab. 4*) suggests that attributive clauses of the form *<N + is a/an ADJ + N>* are prototypical constructions in PLS. This structure is of course used in many other types of discourse, but in this context the pattern introduces either: (1) an 'epithet', that is to say an explicitly evaluative adjective (e.g. *is a + major health problem*, *major risk factor*, *essential trace element*, etc.), or (2) a 'classifier' which specifies a characteristic in a technical item (as in: *is a + chronic disease*, *common disorder*, *long-acting drug*, etc.). There is thus a tendency for authors to provide highly explicit evaluations (as in *common*, *essential* etc.) as well as definitions. *Table 12* gives some examples of this.

Table 12. – Attributive clauses providing definitions and evaluations in PLS.

<VBZ DT JJ NN>
Multiple sclerosis (MS) **is a chronic disease** of the nervous system.
Naltrexone **is a long-acting drug** (an opioid antagonist) [...]
Dry eye **is a common disorder** of the tear film [...]
Selenium **is an essential trace element** [...]

6.3. *Explicitation: presenting numerical data as words*

Perhaps one of the most striking differences between ABS and PLS lies in the presentation of numerical results. Thus, two different results discussed above turn out to be entirely complementary: namely in section 5.2 (see *Tab. 3*) we find that that comparative adjectives and adverbs are overrepresented in PLS, while in 5.7 (see *Tab. 8*) cardinal numbers are seen to be underrepresented in PLS. We suggest that numbers and

figures in ABS are typically reformulated as words in PLS. This process involves various techniques. In many cases, abbreviations belonging to the field of statistics are spelled out (*RR > Risk Ratio*; *HR > Hazard Ratio*). In other cases, abstract or technical quantities are expressed as explicit evaluations (e.g. *50% > good levels*; *RR 5.11 > more likely*) or paraphrased by more 'homely' referents (*100 mg > a mug of coffee*). Additional examples are given in *Table 13*.

Table 13. – *Lexicalised presentation of numerical comparisons.*

Abstract	Plain Language Summary
at least 50% pain relief	**good levels** of pain relief
congestive heart failure: RR 5.11	risk of heart toxicity is **five times more** likely
overall survival: HR 0.66	mortality is reduced **by one-third**
caffeine used at doses of 100 mg or more	a dose **equivalent to a mug of coffee**

Of the different sections in the ABS (which we have not had space to develop in this article), it is the function of the Results sub-section to present all the findings of the statistical tests performed for comparing trials included in the SR in numerical format. While SRs and ABS present these results in some detail, only a few key elements are selected for presentation in the PLS. The example below (see *Ext. 3*) gives an idea of what this looks like in one specific pair of ABS/PLS.

Extract 3. – *Simplified presentation of numerical results.*

Abstract – Main results	Plain Language Summary
In a mixed population trial, qualitative analysis of patients with RA showed a significantly greater reduction in pain favouring topical capsaicin over placebo at one and two weeks (MD -23.80, 95% CI -44.81 to -2.79; NNT 3, 95% CI 2 to 47; MD -34.40, 95% CI -54.66 to -14.14; NNT 2, 95% CI 1.4 to 6 respectively).	In the **one small study** testing capsaicin cream (**0.025%**) in **patients** with persistent knee pain, **patients** also had **better** pain relief with capsaicin cream **than** for those given a placebo cream. **On average, patients** receiving the active treatment improved by **34 more** points (**out of 100**) than the **control group**.

In *Extract 3*, only one numerical result has been selected for presentation in the PLS: this corresponds to the average pain relief provided by the treatment over placebo after two weeks of use. The abbreviation *MD* (*Mean Difference*) has been made explicit lexically (*On average*), as has the comparison (*more*). Also, it can be seen that the presentation has been simplified, from a negative number with decimals in the Abstract (*-34.40*) corresponding to a reduction in pain score, to a positive rounded

number (*34*) in the PLS that corresponds to an improvement on pain assessment scale. Finally, as discussed below, *Extract 3* gives a very clear instance of how human participants (i.e. *patients*) can be expressed as the topic of the main clause (in Subject position) as opposed to the non-animate Subject in the ABS text (*qualitative analysis*).

6.4. *Focus on the human user: putting people first in the clause*

As can be observed in many of the previous examples, human participants are often brought into Theme (Topic) position in PLS. In many cases this involves the word *people* (a 'pantonym' or general word), as well as other plural nouns referring to human participants, such as *patients*, *doctors*, etc. The function of these generic nouns is to replace abstractions such as *numbers* and *levels*. Similar reformulations were noted in section 5.6 (see *Tab. 8*) in which personal pronouns (references to *you*, *your*), but also relative pronouns (such as *who*, *whose*) were found to be key items, as can be seen in *Table 14*.

Table 14. – *Topicalisation of human participants.*

Abstract	Plain Language Summary
Age-related cataract is the opacification of the lens, which occurs as a result of denaturation of lens proteins.	**As people get older**, the lens in the eye can become cloudy – this is known as a cataract.
Conservative management is through bed rest and immobilisation once the acute symptoms have settled.	**People are treated in hospital** either conservatively by being placed in a lying position that reduces strain [...].
Observational studies suggest that antivenom improves outcomes for venom induced consumption coagulopathy in some snakebites and not others.	There is some evidence from observational studies **in humans** which suggest that snake antivenom is helpful to **people who have been bitten by a snake**.

There is thus a general tendency in the PLS to refocus the discourse towards the human user and his/her experience, especially through the frequent occurrence of plural nouns such as *patients*, *people*, *babies*, *women*, *smokers*, etc. as active participants in processes. This observation also correlates with a general preference to re-express medical phenomena in terms of the concrete localisation of processes in the human body.

It is important to emphasise here that the topicalisation of humans has an impact on various other grammatical systems. Thus the examples above show how the Theme of the discourse (in Halliday's terms this is the first item in the clause that is typically expressed as Subject) shifts

from ABS to PLS. In ABS the typical Themes involve non-animate conditions (e.g. *age-related cataract*), indirect causes of disease (*snakebites*) or treatments (*conservative management*), while in PLS the typical Themes become animate *people* (this item being a highly significant keyword). Such a reformulation necessarily involves other grammatical changes (shift from passive to active, replacement of embedded adjectives or other specific items by clause-modifying subordinate clauses, post-modifying relative clauses, etc). It is thus notable how the so-called 'simplified' discourse becomes more complex (or as mentioned above, more 'intricate') in terms of syntactic and morphological structure. For example, it can be seen that the simple metonymic use of *snakebite* in ABS is re-conceptualised by various defining clauses in PLS: *people who have been bitten by a snake*, *people who have a snakebite*, etc.

A final point on the 'humanisation' of PLS: as mentioned above, our characteristic elements analysis shows that personal pronouns are over-represented (see *Fig. 1* and *Tab. 8*) in PLS. We suggest that this is related to directly addressing the reader in an imaginary dialogue. To give just one example: *Usually,* **you** *connect two electrodes from the machine to* **your** *skin on the painful area.* **Your** *doctor or physiotherapist will show* **you** *how to use it, and most machines can be used at home.*

6.5. *Focus on the human user: making generalisations and giving clear judgments*

Lastly, the overrepresentation of present tenses and modal verbs in complex verbal phrases in the PLS (sections 5.3, 5.4 and 5.5, see *Tabs. 4, 5* and *6*) can be associated with subtle differences in expression, which we interpret as both generalisation and more explicit interpretation of the results in the SR. *Extract 4* below represents a typical example of this.

Extract 4. – Generalisation and interpretation.

Abstract – Main results	Plain Language Summary
No separate safety data were available for patients with RA, however 44% of patients developed burning at the site of application and 2% withdrew because of this.	About 50 % of **patients who** use capsaicin cream on their skin will develop this local burning but **only** 2 in 100 **will stop** treatment because of this.

This example shows several features relating to generalisation, or rather the shift from the specific environment of clinical research findings to a

more 'mediated' discourse, which relates general tendencies for the non-expert reader. Thus we can see in this example: (1) the shift from the past tense of specific clinical trials in the ABS (e.g. *developed*, *withdrew*) to the present tense of general truth of findings in the PLS (e.g. *develop*, *stop*), (2) the shift from relational processes (using relational verbs) to expressions from the patient's perspective (with dynamic, user-oriented verbs) modified by modal auxiliaries (e.g. *will*) and other markers of modality (e.g. *about*, *but only*). Thus, while the ABS aims for a factual account of trial results without any qualitative interpretation – in accordance with Cochrane guidelines, which instruct authors to avoid making recommendations – the PLS offers something of a generalised conclusion, as interpreted from the user's perspective.

7. Conclusion: some recommendations

In this paper we have attempted to identify some of the characteristic features of 'plain language' as it is used spontaneously by authors of Cochrane Plain Language Summaries. Using textometric analysis we have identified some of the most outstanding features of PLS as they compare with texts that have the same context, but are aimed at a completely different audience (Systematic Review Abstracts – ABS).

Our first observation is that while ABS typically use compact specialised terms and impersonal expressions, which clearly belong to the domains of medicine and statistics, such structures are broken down in PLS, which typically use post-modifying prepositional phrases and relative clauses, and frequent inclusion of definitions. Also, while ABS typically express comparisons of interventions in numerical format, in PLS the same comparisons are reconceptualised as 'verbal' (i.e. non-numeric) descriptions of quantities and abbreviations as well as lexical markers of comparison.

In translation studies, these techniques are known as 'explicitation'. We find that they are also consistently combined with writing practices that re-centre or focus the discourse on the human user and his/her specific experience. This includes strategies such as the topicalisation of human participants as well as the concrete localisation of processes in the human body. Other indices of 'user-orientation' in the PLS include frequent use of interactive discourse features which address the reader directly, the generalisation of trial results through the use of the present tense (as opposed to the past of reporting in ABS) and the explicit sig-

nalling of authorial stance (explicit statement of results) accompanied by the use of modal markers.

We can present our findings as a short list of general recommendations:

• Break down complex information: (a) unpack compact nominals (for details see section 6.1), (b) use simple attributive clauses as definitions, including the active and the passive (6.2), (c) present numerical data as words (6.3).

• Focus on the human user: (a) put people first in the clause (6.4), (b) make generalisations and give clear judgments (6.5).

Some of these recommendations look like the general advice already given in the Cochrane style guide and elsewhere (for example "make generalisations"). However, the difference is that we couch our recommendations in linguistic terms, and in section 6 we set out specific patterns (presented as lexico-grammatical sequences) as well as examples which can be used as models (only space precludes us from repeating them here). Also, it is interesting to note that, in some cases, our recommendations go against Cochrane's advice. For example, the Cochrane style guide currently advises authors to avoid the passive (as do many other plain language guides). But our data analysis suggests that the passive is a key construction in PLS, where it is used to summarise results and recommendations efficiently without going into technical detail (see section 5.4 and 5.5 for a discussion of this).

Of course, we cannot set out such a short set of recommendations without also raising a number of difficulties. In the first instance, not all linguists or technical writers will agree on how to interpret our metalanguage. In addition, recommendations such as "make generalisations…" clearly require more specification. Such objections raise questions about who such recommendations are aimed at, and why they need to be formulated. It is therefore perhaps worth emphasising here that over the years Cochrane has found it necessary to produce increasingly explicit guidelines for authors of SRs (and therefore ABS and PLS 2). However, we would suggest that since these guidelines are not based on corpus analysis, it is still useful to identify in a systematic manner the most salient features of PLS in order to present these in a list of prioritised features.

---

[2] Most recently, in 2017, a pilot project conducted by Cochrane Norway set out guidelines for authors of PLS that recommend strategies corresponding to some of the tendencies observed in the present study (e.g. using questions as titles and signalling level of certainty through modal markers).

Finally, the study presented in this paper only represents a preliminary step in a much broader series of related research projects. For example, many ABS/PLS pairs happen to have different titles, and our preliminary findings suggest that the lexico-grammatical differences between titles of ABS and PLS are not quite the same as those observed in the main bodies of the ABS and PLS. Thus there are many interesting aspects of SR, ABS and PLS which remain to be explored. More generally, there are also normative questions about the quality of the texts to be analysed (should we be studying all the PLS as a single corpus, or should we be studying a subset of 'exemplary' PLS versus less successful texts?). And a much broader further area of study involves text ergonomics and training. For example, how should we formulate our recommendations for Cochrane authors (or technical editors)? How much linguistic meta-knowledge can we assume? And in the long term, what will be the relevance of linguistic guidelines in a context where authors will increasingly use on-line platforms or computer-assisted authoring packages? Naturally, these questions fall outside the remit of the present paper, but they also underline the importance of the first step in our research programme: describing how authors simplify their texts on the basis of the systematic corpus-based analysis of linguistic regularities.

References

Blum-Kulka, S. 1986. "Shifts of Cohesion and Coherence in Translation". In *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*, edited by Juliane House and Shoshana Blum-Kulka, 17-35. Tübingen: Gunter Narr.

Chervak, Steven, Colin G. Drury, and Jean-Pierre Ouellette. 1996. "Field Evaluation of Simplified English for Aircraft Workcards". In *Proceedings of the 10th FAA/AAM Meeting on Human Factors in Aviation Maintenance and Inspection*, Alexandria (VA), January 1996. [22/07/2019]. https://doc.uments.com/s-field-evaluation-of-simplified-english-for-aircraft-workcards.pdf.

Cochrane Style Manual. [22/07/2019]. https://community.cochrane.org/style-manual.

Crossley, Scott, Max Louwerse, Philip McCarthy, and Danielle McNamara. 2007. "A Linguistic Analysis of Simplified and Authentic Texts". *The Modern Language Journal* 91 (1): 15-30.

Cutts, Marin. 2013. *Oxford Guide to Plain English*. Oxford: Oxford University Press.

Evert, Stefan, and Andrew Hardie. 2011. "Twenty-first Century Corpus Workbench: Updating a Query Architecture for the New Millennium". In *Proceedings*

of the Corpus Linguistics 2011 Conference*, University of Birmingham, July 20-22, 2011. [22/07/2019]. https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf.

Farrington, Gordon. 1996. "AECMA Simplified English: An Overview of the International Aircraft Maintenance Language". In *Proceedings of the First International Workshop on Controlled Language Applications (CLAW96)*, March 26-27, 1996, 1-23. Leuven: Katholieke Universiteit Leuven Centre for Computational Linguistics.

Fleury, Serge, and Maria Zimina. 2014. "Trameur: A Framework for Annotated Text Corpora Exploration". In *System Demonstrations*. Proceedings of COLING 2014 (25th International Conference on Computational Linguistics), Dublin, 57-61. [22/07/2019]. https://aclanthology.coli.uni-saarland.de/papers/C14-2013/c14-2013.

Fonseca, David. A. 2006. "How Simple is Simplified Technical English?". *Intercom* 53 (2): 20-22.

Gledhill, Christopher, Stéphane Patin, et Maria Zimina. 2017. "Lexico-grammaire et textométrie. Identification et visualisation de schémas lexico-grammaticaux caractéristiques dans deux corpus juridiques comparables en français". *Corpus* 17: 113-144. [22/07/2019]. https://journals.openedition.org/corpus/2726.

Grimshaw, Jeremy. 2004. "So What Has the Cochrane Collaboration Ever Done for Us? A Report Card on the First 10 Years". *Canadian Medical Association Journal* 171 (7): 747-749. [22/07/2019]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC517860/.

Halliday, Michael A.K., and Christian M.M. Matthiessen. 2014. *Introduction to Functional Grammar*. 3rd ed., 4th ed. London: Edward Arnold.

Higgins, Julian P.T., and Sally Green, eds. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0 (updated March 2011). The Cochrane Collaboration. [22/07/2019]. http://handbook-5-1.cochrane.org/.

Lafon, Pierre. 1980. "Sur la variabilité de la fréquence des formes dans un corpus". *Mots* 1 (1): 127-165.

Lebart, Ludovik, André Salem, and Lisette Berry. 1991. "Recent Developments in the Statistical Processing of Textual Data". *Applied Stochastic Models and Data Analysis* 7 (1): 47-62.

Lebart, Ludovik, André Salem, and Lisette Berry. 1998. *Exploring Textual Data*. Dordrecht: Kluwer Academic Publishers.

McIlwain, Catherine, Nancy Santesso, Silvana Simi, Mary Ann Napoli, Toby Lasserson, and Emma Welsh. 2013. *Standards for the Reporting of Plain Language Summaries in New Cochrane Intervention Reviews (PLEACS)*. [22/07/2019]. https://methods.cochrane.org/sites/default/files/public/uploads/PLEACS_0.pdf.

O'Brien, Sharon. 2003. "Controlling Controlled English: An Analysis of Several Controlled Language Rule Sets". *Machine Translation Archive*. [22/07/2019]. http://www.mt-archive.info/CLT-2003-Obrien.pdf.

Pinker, Steven. 2014. *The Sense of Style*. London: Penguin.

Renahy, Julie, Dominique Vuitton, Barbara Rath, Izabella Thomas, Valérie de Grivel, and Sylviane Cardey. 2015. "Controlled Language and Information on Vaccines: Application to Package Inserts". *Current Drug Safety* 10: 41-48.

Schmid, Helmut. 1994. "Probabilistic Part-of-speech Tagging Using Decision Tree". In *Proceedings of the International Conference of New Methods in Language Processing*, Manchester, vol. 12, 44-49. [22/07/2019]. http://www.cis.uni-muenchen.de/-schmid/tools/TreeTagger/data/tree-tagger1.pdf.

Stewart, Kathleen M. 1998. *Effect of AECMA Simplified English on the Comprehension of Aircraft Maintenance Procedures by Non-native English Speakers*. Masters Dissertation, University of British Columbia, Department of Reading Education. [22/07/2019]. https://open.library.ubc.ca/media/download/pdf/831/1.../2.