**Research Articles From *Plos Biology*: A Textual Data Analysis**

Bruno Oberle, LiLPa, University of Strasbourg (France)
oberleb@unistra.fr

**Abstract:** Many research articles in experimental sciences present a standardized form known as the "IMRaD format", an acronym for *Introduction, Methods, Results and Discussion*. Linguists have pointed out that each section serves a specific rhetorical function. This article aims at exploring the relation between linguistic features and the rhetorical function of each section. We studied 966 research articles from the journal *Plos Biology*. Topic modelling and correspondence analysis highlighted linguistic features associated with the rhetorical function of each section. However, lexical diversity revealed some features that point out a change in the rhetorical function traditionally associated with *Methods* sections, because *Plos Biology* uses a slightly modified IMRaD format, in which *Methods* sections are at the end of articles.
**Keywords:** academic writing, research article, IMRaD, corpus analysis, topic modelling

## 1. Introduction

Many research articles in experimental sciences present a standardized structure known as the "IMRaD format" (Bazerman 1988). The acronym stands for *Introduction*, *Methods*, *Results and Discussion*, which refer to the main sections of the article.

This format is usually imposed by journal style sheets or recommended by manuals of style (like the *Publication Manual* of the American Psychological Association).

Each section has a particular rhetorical function (Müller-Gjesdal 2013): the *Introduction* presents the research question and relates it to previous works; the *Methods* section describes the data and procedures used for the experiment; the *Results* section reports the observations and the *Discussion* section tries to put them into perspective in the light of the theoretical framework selected in the introduction.

These rhetorical functions induce a routinization of the writing process, since "each section must conform to detailed instructions, at times resembling a questionnaire in specificity" (Bazermann 1988, see also Rinck 2010). It is this routinization that allows the sections to be contrasted, "*e.g.* Introductions in contrast to Methods" (Swales 1990), since different sections are characterized by different linguistic features. For example, Reimerink (2006) has shown that the semantic type of verbs varies according to the section, Müller-Gjesdal (2013) that the French pronoun *on* is not uniformly distributed across medical research articles, while Bertin and Atanassova (2014) have studied the lexical distribution of citations in relation to the different IMRaD sections.

In this context, our research aims to characterize the four IMRaD sections by using the textual data analysis methods, in a corpus of 966 articles from the journal *Plos Biology*.

We will first focus on lexical frequency and study the diversity of terms in different IMRaD sections. We will then consider the articles from the point of view of topic analysis to find what are the most prominent topics in each section. Finally, we will study parts of speech using a correspondence analysis.

## 2. The corpus

We collected 966 research articles in English from *Plos Biology*, a journal in open access[1]. We automatically retrieved 1 090 articles from January 2010 to October 2016, but removed articles that did not have the four sections we wanted to analyze (in some articles, for example, the *Results* and *Discussion* sections are not separated). We thus have 3 864 sections (four sections for each of the 966 articles).

[1] http://journals.plos.org/plosbiology

Articles were available in XML format, and sections were identified by specific tags. XML attributes clearly indicated section types for the most recent years. For older articles, we had to infer this information from the title of the section.

The content of the articles was directly available in the XML file in a standardized form, and we did not need any further preprocessing. End notes were not included, but figure and table captions (sometimes very long) were kept when they were inside a section (we have however discarded all the tables and images added as complements at the end of the article).

The text of the 3 864 sections was tagged with part of speech information with the software TreeTagger (Schmid 1994, Schmid 1995). The Table 1 shows the number of tokens (that is, the number of occurrences of each item, including punctuation), types (the number of distinct items) and lemmata.

| | **tokens** | **tokens (average)** | **types** | **lemmata** |
|---|---|---|---|---|
| **results** | 5 097 672 | 5 277 | 104 232 | 122 439 |
| **methods** | 2 053 524 | 2 125 | 89 913 | 97 264 |
| **discussion** | 1 669 997 | 1 728 | 52 604 | 59 792 |
| **introduction** | 887 008 | 918 | 39 285 | 43 471 |
| **corpus** | 9 708 201 | 2 512 | 175 405 | 203 442 |

*Table 1: Tokens, types and lemmata for the 966 articles (3 864 sections), in descending order of number of tokens.*

There is a large difference in size between the sections. *Results* sections are usually the longest, with an average of 5 277 tokens, almost five times more than the *Introductions*, the shortest sections. This fact must be taken into account in textual analysis, since some computations are sensitive to the length of the texts compared.

## 3. Lexical analysis

We begin our lexicometric analysis by the study of word frequencies, focusing on lexical diversity. From the rhetorical functions presented in the introduction, we can hypothesize that *Introductions* and *Discussions* should have a greater lexical diversity, since they are related to other works. On the contrary, *Methods* sections should have the lowest lexical diversity, since they are more standardized and more focused on the experiment.

To test this hypothesis, we sorted sections by lexical diversity. Because this measure is length sensitive, we took only samples[2] of the same length (based on the shortest section).

For most of the articles (about 67%), *Methods* sections have the greatest lexical diversity (see Table 2). On the contrary, *Results* sections are the least diverse.

[2] Note that the samples are composed of non-consecutive tokens. We also compared the results with samples of consecutive tokens. While the percentages vary somewhat (the proportion of articles in which *Methods* sections are the most diverse is around 45%), the conclusions remain the same (there are more articles for which *Methods* are the most diverse). We also tested with a continuous set of 500 tokens, for all parts and all texts: *Methods* are again the most diverse. Finally, we tried to use lemmata rather than words, but the difference is not significant.

| | nb of articles | percentage | average diversity |
|---|---|---|---|
| **methods** | 642 | 67.15% | 0.45 |
| **discussion** | 138 | 14.44% | 0.42 |
| **introduction** | 96 | 10.04% | 0.41 |
| **results** | 80 | 8.37% | 0.41 |
| **total** | 956 | 100% | |

*Table 2: Results of the lexical diversity analysis. Since the samples are selected at random, repetitions of the experiment may yield slightly different results each time. The column "number of articles" shows the number of articles for which the indicated section is the most diverse. Ten articles had to be removed because some of their sections were too short, usually a text pointing to a website, for example: "The materials and methods used for this report can be accessed online." (1000543)*

Our hypothesis is *not* confirmed by these results, and we must look at the texts themselves to understand why. We first need to note that the articles from our corpus do not follow the usual IMRaD order (only eleven articles out of 966 keep this order), since *Methods* sections are placed at the end of the paper, and not between *Introductions* and *Results*.

This is why *Methods* sections seems to have a slightly different role. Rather than presenting the experiment and how it was conducted, it is rather a list of tools, instruments and materials. This list is technical and often specifies models, brands, manufacturers, *etc.*:

(1) Sony Vegas Pro 8.0 (Sony, Tokyo, Japan) (1002534[3])

(2) Proteins were immobilized on magnetic glutathione beads (Pierce Biotechnologies Inc., Rockford, IL) (1002534)

or the type of the microbial strains used in the experiment:

(3) All yeast strains were isogenic to AM1003 [23], which is a chromosome III disome with the following genotype: hmlΔ::ADE1/hmlΔ::ADE3 MATa-LEU2-tel/MATα-inc hmrΔ::HYG FS2Δ::NAT/FS2 leu2/leu2-3,112 thr4 ura3-52 ade3::GAL::HO ade1 met13. (1000594)

There may also be a detailed list of techniques and softwares:

(4) Statistical differences between datasets were analyzed with two-tailed unpaired Student's t tests from which p-values were derived. Scatterplots were generated using Graphpad Prism 5.0 (Graphpad Software, La Jolla, CA) (1002534)

or a description of very specific protocols:

(5) Beads were washed 6 times with binding buffer, and fusion proteins were eluted after a 30 min incubation in elution buffer (125 mM Tris pH 8.0, 150 mM NaCl, 50 mM glutathione) containing HALT protease inhibitor (Pierce Biotechnologies Inc. Rockford, IL) at 4°C (1002534)

This accumulation of technical details explains why *Methods* sections have a high lexical diversity. The description of the experiment itself is rather done in *Results* sections:

(6) To determine if sudden collapses in the intestinal Aeromonas abundance occur in the absence of Vibrio, *we examined* live imaging data of Aeromonas mono-associations over a similar time frame... *We next inspected* the spatial structure and

[3] Articles are identified by a number extracted from the DOI. To access the article, just go to this address: http://dx.doi.org/10.1371/journal.pbio.<ID> where <ID> is the number given at the end of the quotation.

dynamics of each species to uncover clues regarding possible factors driving Aeromonas' sudden population drops (1002517)

*Methods* sections merely lists references to other articles, without explanation, as shown in the following example (the numbers between brackets are references given at the end of the *Plos Biology* article):

(7) Wild-type AB or ret mutant (ret1hu2846, ZFIN ID: ZDB-ALT-070315-12) zebrafish were derived GF and colonized with bacterial strains as previously described [19].

Aeromonas (ZOR0001, PRJNA205571) and Vibrio (ZWU0020, PRJNA205585) were isolated from the zebrafish intestinal tract as described previously [24].

Dissection of larval guts was done as described previously [19].

Imaging was performed using a home-built light sheet fluorescence microscope, based on the design of Keller et al. [18] and described in detail elsewhere [21,22].

The analysis pipeline used to estimate bacterial abundances from light sheet imaging is described in [21].

Sample mounting is done as previously described [21].

etc. (1002517)

Here we can see that most of the paragraphs of *Methods* sections begin with an element or a protocol "described elsewhere" (with a reference to another article). *Methods* sections are thus lists, not descriptions as we would expect for more traditional IMRaD articles (that is, articles with the *Methods* section right after the *Introduction*).

It should be noted, however, that this is only a trend: some articles have more narrative *Methods* sections (see the article 1002564 for an example).

The study of lexical diversity has thus shown that *Methods* sections have the most diverse vocabulary, contrary to what we had expected. The place of this section explains this fact: at the end of the article, *Methods* somewhat lose their original rhetorical function (description of the experiment) and are reduced to a mere list of tools and materials.

## 4. Topics

Sections can also be characterized by their most prominent topics. We can expect two types of topics: domains of biology, which may be found in all sections, and topics more specific to each section (for example *Methods* sections may have a topic for tools and instruments).

For this study, we used TopicModellingTool-Fr (Hengchen 2015), based on the Mallet tool (McCallum 2002). In order to limit irrelevant results from units of measure and various symbols extremely frequent in biology texts, we only considered tokens of more than four characters. Moreover, we applied the tool both on lemmata and words: the results are very similar. We will only give here the results for the analysis on words.

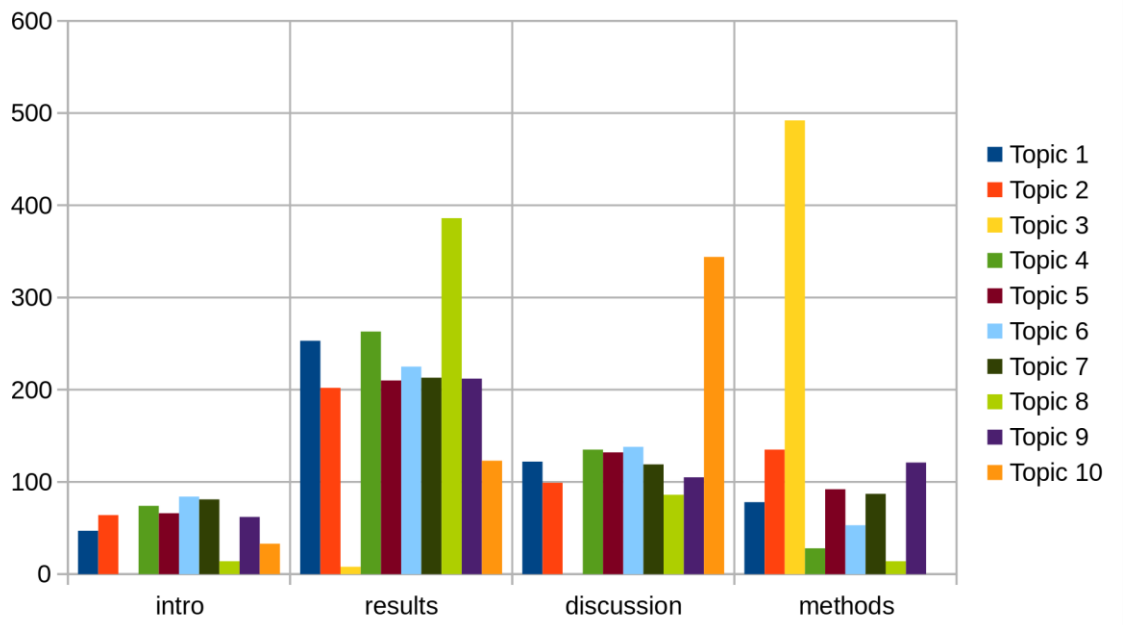


*Figure 1: Distribution of documents according to sections and topics (10 topics).*

From the whole corpus (966 documents, 3 864 sections), we let the tool find 10 topics. The Figure 1 shows the distribution of the documents by sections and topics. While, most of the topics appear in all sections (the highest frequency of documents from *Results* sections comes from the fact that these sections are longer than the others, so it is not surprising), three of them (Topics 3, 8 and 10) seem to be overrepresented in one section only (respectively *Methods*, *Results* and *Discussions*), and almost absent from the other three sections. Topic 3, for example, is overrepresented in *Methods*, and does not appear in *Introductions*, *Discussions*, or *Results*. These topics are composed of specific terms, for example the vocabulary of material processing (*incubated, washed, performed, etc.*) for Topic 3.

Nevertheless, perhaps because of the diversity and precision of research articles, reading the other topics is not easy. Therefore we have gradually increased the number of topics so that they could be more easily isolated. We stopped at 22 topics. This may seem a lot, but this covers all the fields of biology and allow us to identify specific topics for most of the sections (see Figure 2).

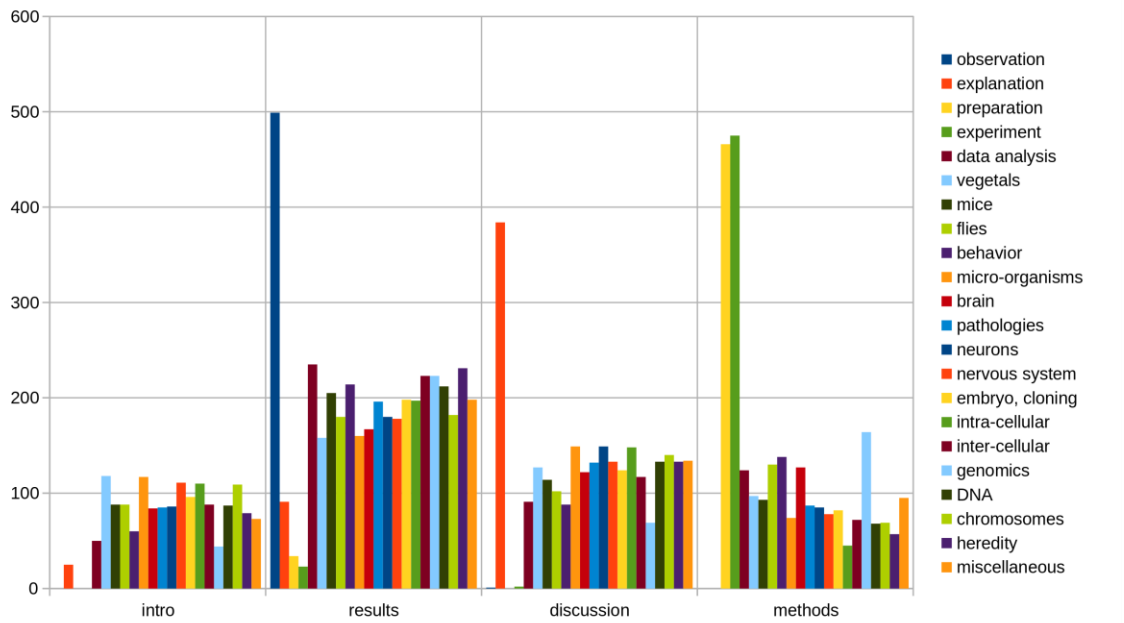


*Figure 2: Distribution of documents according to sections and topics (22 topics).*

*Methods* sections have two specific topics, one oriented towards the preparation of the materials used in the experiment (*incubated, performed, washed, purified, assay, sample, analysis, etc.*), the other oriented towards the tools (*performed, experiments, microscope, software, image, calculated*). *Results* sections are characterized by the vocabulary of observation and analysis (*observed, test, induced, analysis, results, control, etc.*), and *Discussion* sections by the vocabulary of explanation (*study(ies), function, mechanism, suggest, evidence, factor, involved, etc.*). These terms are related to the rhetorical function of each section. However, we can be surprised at the lack of specific terms for *Introductions*: we could have expected topics similar to those of *Discussions*.

The other topics are evenly distributed among the sections (in proportion to the size of each section). They relate to the fields of biology. We thus find animals used in the experiments (*mice, flies, plants, micro-organisms*), research fields of biology (*studies on the brain, behavior, pathologies, neurons, the nervous system, embryos, cloning, biology intracellular and inter-cellular*), DNA studies (*genome, chromosome, heredity*), *etc*.

To conclude this section, we should remember the presence of specific terms for three of the four IMRaD sections, and the fact that these terms may be found with topic modeling tools.

## 5. Parts of speech

In the previous sections, we have characterized the IMRaD sections by analyzing their lexical diversity and their most prominent topics. We will now focus on how similar or different these sections are. Since both *Introductions* and *Discussions* are related to other works, we may hypothesize that these two sections are similar. *Methods* and *Results* sections, on the other hand, are expected to stand apart from each other.

We will explore this question using a correspondence analysis, computed with the software TXM (Heiden 2010). We used as variable the parts of the speech, tagged with TreeTagger (Schmid 1994, Schmid 1995), because they allow us to highlight more linguistic aspects than mere words or lemmata.

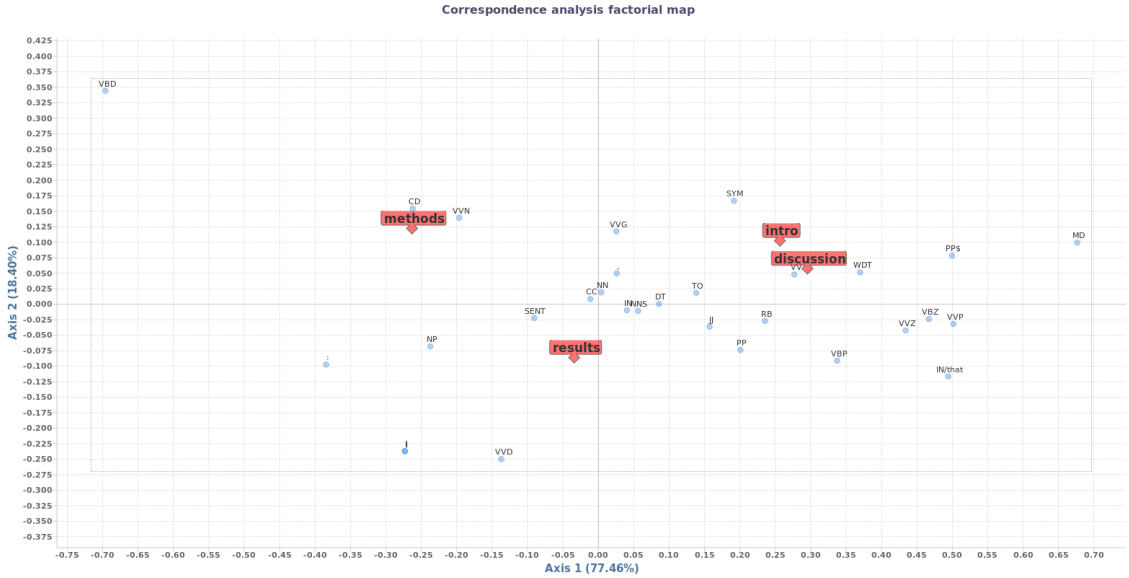


*Figure 3: Correspondence analysis factorial map of parts of speech. Only the 20 most frequent parts of speech are represented. The tagset used is the one of the default model of TreeTagger.*

Figure 3 presents a factorial map that shows the proximity of *Introductions* and *Discussions*, two sections that differ significantly from *Results* on one hand, and *Methods* on the other hand. To understand the relationship between a section and a part of the speech, we have to look at the angle at the center between dots representing sections and dots representing parts of speech (Cibois 2015 p. 31). Three scenarios are possible: if the angle is acute, then there is an "attraction" between the modalities (sections and parts of speech); if it is obtuse, there is on the contrary an "opposition"; and if it is close to 90° then there is independence (and the relation is not interesting) (Cibois 1987, Cibois 2015). It is therefore the angle that indicates the attraction or opposition between the modalities, and not only the proximity, especially for dots close to the center.

*Introduction* and *Discussion* sections are close together. This means that they are attracted and opposed to the same parts of speech. Verbal bases (tagged as "VV") seem to correlate with these parts, but they are one of the most problematic part of speech, since there are a lot of errors in the tagging. So we will analyze other parts of speech.

Possessive pronouns (tagged as "PP$") appear most often in *Introductions* and *Discussions*: authors use them to relate their own research to other works:

(8) A mechanism to explain this was described by C.H. Waddington in *his* influential book "Organizers and Genes" (1001907)

Modals (*could, should, will, etc.*, tagged as "MD") also appear in *Introductions* and *Discussions*. This is not surprising, since these are the only sections that allow modalization of discourse, whereas *Methods* and *Results* sections are expected to be more objective and neutral in tone. In *Introductions*, modals are used to make assumptions:

(9) If so, then it *should be possible* to classify... Successful classification *would validate* multivariate decoding of unconstrained brain activity... (2000106)

while in *Discussions*, they are more often used to generalize an observation, with caution:

(10) E. editha does not mate on its host on a micro-scale... This means that neither immigrant inviability nor habitat isolation *should prevent* local insects from mating with differently adapted migrants. (1000529)

or to give explanations:

(11) This *could explain* the finding that microglial motility decreased during light deprivation. (1000527)

The "wh-determiners" (tagged as "WDT"), that is, relative pronouns that have an adverbial for antecedent, especially *that* and *which*, are also mostly found in *Introductions* and *Discussions*. This may be correlated to a greater length of the sentences of these sections (they are five words longer than the sentences of other sections), but this is mostly the result of a larger presence of explanations:

(12) Individuals affected by the Lynch syndrome undergo somatic inactivation of the second allele *that causes the impariment of the MMR machinery*... The genetic condition is known as hereditary non-polyposis colorectal cancer (HNPCC), *which represents the most common form of inherited colorectal cancer*. A hallmark of MMR deficiency is microsatellite instability (MSI), *which measures the accumulation of insertions and deletions (indels) at repeated regions of the genome*. (1000275)

We might be surprised that there are no proper names (tagged as "NP") in *Introductions* and *Discussions*, since it is in these sections that authors relate their research to other works by other authors. There are two explanations for this. The first concerns the tagging done by TreeTagger: tokens tagged as proper names are in fact not proper names, but technical terms, acronyms and units of measure, like *RNA*, *Drosophilia*, *NaCl*, *ANOVA*, *etc*. Strangely, the most common "proper name" (according to TreeTagger) is *Figure*, while *Darwin* (who appears 13 times in the corpus), is never tagged as such. However, and this is the second explanation, the texts actually contain few names of researchers, since the references are given systematically by a number between brackets: the authors appear in the notes, not in the text.

*Methods* sections are strongly correlated to cardinals (tagged as "CD"). As we have said before, *Methods* are very precise and give a lot of quantified indications in protocols (*for all reactions, µl of cDNA was used in a µl qRT-PCR reaction* (1002499)), but also mathematical values (*FIMO was run using the search criterion of p <0.0001*), software versions (*we used BLAST (v.2.2.28 +) to identify ... homologous sequences in D. yakuba*), *etc*.

Past forms of *be* (*was, were*, tagged as "VBD") and past participles (tagged as "VVN") are also overrepresented in *Methods* sections. This is due to the use of passive forms in descriptions:

(13) All protocols *were reviewed* and approved... Postnatal day (P) 17--25 gerbils *were used* to generate thick (450--500 µm) horizontal slices... Animals *were deeply anesthetized*... The brain *was then dissected* free in 32°C oxygenated ACSF, and one horizontal slice *was obtained* with a Leica vibratome... (1000406)

On the contrary, *Results* sections use more active past forms (tagged as "VVD"); authors describe here how results were obtained, and this narration is in the active voice:

(14) We *plotted* the curvature x disparity interaction effect... In every animal, we *observed* a decrease... Note, however, that monkey S *showed* a significant effect of CIP inactivation... (1002445)

This is probably different from more traditional IMRaD articles, in which narrations of this kind are in *Methods* sections. Here, we begin to see a shift in rhetorical functions of *Methods* and *Results* sections.

The shift can also be noticed with personal pronouns (tagged as "PP"), which are under-represented in *Methods* sections: this confirms the use of passive forms, but is surprising, since we would expect a larger use of first-person pronouns (*I, we*) in the description of the methodology. Again, this is explained by the variant of the IMRaD format used in *Plos Biology*, in which *Methods* sections are at the end of articles: *Methods* are no longer a narrative of what has been done, but rather a list of what has been used. The narrative is most often found in

*Results*, and this is why first-person pronouns appear more often in *Results* than in *Methods* (0.28% *versus* 0.17%).

The correspondence analysis has shown that IMRaD sections can be divided into three groups: results, methods, and introductions/discussions. This is globally in accordance with our hypothesis. The linguistic features of *Methods* sections, however, confirm the shift in rhetorical function that we evoked at the beginning of this paper.

## 6. Conclusion

Our goal was to characterize IMRaD research articles with textual data analysis methods. Our study has shown that each section has its own linguistic features, and this confirms the more qualitative analysis of previous works. The topic analysis, for example, has shown that each section has its own vocabulary, while the correspondence analysis has shown that *Results* and *Methods* sections do not share the same parts of speech.

However, our hypotheses, built on the rhetorical functions we found in previous works (most notably Swales 1990, Swales 2004 and Gjesdal-Müller 2013), have not all been confirmed. The study of lexical diversity, which is often more important in *Methods* sections, has for example shown that articles from *Plos Biology* use a variant of the IMRaD format, in which *Methods* sections are at the end of the article, and not right after *Introductions*.

Our study therefore raises new questions about this variant of the IMRaD format, which seems to have been ignored by the literature so far. This variant should be better characterized, especially since the change in order seems to have consequences on the rhetorical functions of some sections, and thus on their linguistic features.

## 7. References


Bazerman C. (1988). *Shaping written knowledge: the genre and activity of the experimental article in science*. University of Wisconsin Press.

Bertin M. & Atanassova I. (2014). A study of lexical distribution in citation contexts through the IMRaD standard. *In Proceedings of the First Workshop on Bibliometric-Enhanced Information Retrieval co-located with 36th European Conference on Information Retrieval (ECIR 2014)*, p. 5-12.

Cibois P. (1988). *L'analyse factorielle.* Presses Universitaires de France.

Cibois P. (2015). *Les méthodes d'analyse d'enquêtes.* ENS Éditions.

Gjesdal-Müller A. (2013). The Influence of Genre Constraints on Author Representation in Medical Research Articles. The French Indefinite Pronoun *On* in IMRaD Research Articles. *Discours. Revue de linguistique, psycholinguistique et informatique*, 12.

Heiden S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. *In* Ryo Otoguro, et al. (eds), *24th Pacific Asia Conference on Language, Information and Computation*, 389-398

Hengchen S. (2015). *Topic Modeling Tool (fr): Working prototype*.

McCallum A. K. (2002). *Mallet: A Machine Learning for Language Toolkit.*

Reimerink A. (2006). The Use of Verbs in Research Articles: Corpus Analysis for Scientific Writing and Translation. *New Voices in Translation Studies*, 2, 9-27.

Rinck F. (2010). L'analyse linguistique des enjeux de connaissance dans le discours scientifique: un état des lieux. *Revue d'anthropologie des connaissances*, 4(3), 427-450.

Schmid H. (1994). Probabilistic Part-Of-Speech Tagging Using Decision Trees. *In Proceedings of International Conference on New Methods in Language Processing*.

Schmid H. (1995). Improvements in part-of-speech tagging with an application to german. *In Proceedings of the ACL SIGDAT-Workshop*.

Swales J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press.

Swales J. (2004). *Research Genres: Explorations and Applications*. Cambridge University Press.