Contribution of Move Structure to Automatic Genre Identification: an Annotated Corpus of French Tourism Websites

Rémi Cardon[1], Trang Pham Tran Hanh[1, 2] Julien Zakhia Doueihi[1], Thomas François[1]
[1] UCLouvain, [2] Hanoi University
remi.cardon@uclouvain.be, tran.pham@uclouvain.be,
julien.zakhia@uclouvain.be, thomas.francois@uclouvain.be

Abstract

The present work studies the contribution of move structure to automatic genre identification. This concept - well known in other branches of genre analysis - seems to have little application in natural language processing. We describe how we collect a corpus of websites in French related to tourism and annotate it with move structure. We conduct experiments on automatic genre identification with our corpus. Our results show that our approach for informing a model with move structure can increase its performance for automatic genre identification, and reduce the need for annotated data and computational power.

Keywords: Web genre analysis, Automatic genre identification, Move structure

1. Introduction

In this paper, we aim to incorporate the concept of "move structure" into a neural architecture for genre identification. This concept, commonly employed in corpus linguistics and English for specific purposes (ESP), serves two crucial functions: (1) it helps to determine the rhetorical organization of a genre to achieve a set of communicative (contextual) goals, and (2) show how those communicative goals are signaled by lexical and syntactic choices. Indeed, move structure assists researchers in clarifying not only lexical and syntactic elements but also the overall structure of discourse.

We propose to explore the role move structure can play in the task of automatic genre identification. As we discuss in section 2, this approach has been overlooked by the community, while representing an important trend in linguistic studies in genre analysis. As we did not find references that would explain why move structure has not been explored, we propose to perform experiments to investigate its potential. Our hypothesis is that informing recent NLP methods (i.e. neural architectures) with move structure may help in performing automatic genre identification. In order to test for this hypothesis, we collected a corpus that we annotated, and conducted experiments with it. The outline of the paper is as follows: we describe what move structure is, along with the existing works for automatic genre identification and automatic genre identification of web genre (Section 2). Then we introduce in detail the corpus that we make available with this work[1] (Section 3). We detail the experiments that we conducted and comment upon the results (Section 4). We finally discuss our insights on the broader theme of web genre definition (Section 6) and conclude (Section 7).

[1] https://github.com/remicardon/genre_moves

2. Related work

2.1. Move Structure

Genre analysis is a framework for the study of specialized communication in a variety of contexts (Adam, 1997, 1999; Bhatia, 1993a; Maingueneau, 2004b,a, 2007, 2016; Swales, 1990; Charaudeau, 2011; Bronckart and Dolz, 2002; Chartrand et al., 2015; Beacco, 2013; Richer, 2011). Genre analysis describes discourse in textual and social contexts, as *"genre is a class of communicative events: the main criterion that turns a collection of communicative events into a genre is represented by some shared set of communicative purposes"* (Swales, 1990, p. 45). Several automatic genre classification studies have been carried out, especially since 2000 (Karlgren and Cutting, 1994; Stamatatos et al., 2000; Dewdney et al., 2001; Lee and Myaeng, 2002; Beaudouin et al., 2002; Santini, 2007, 2006) (see Section 2.2 for more details). In recent years, the preferred properties for genre classification include parts of speech and syntactic features (syntactic functions, numbers of complex nominal groups, subject or object types) Todirascu (2019). However, from our perspective, if we adhere to the strict definition of genre put forth by Swales (1990), it is essential to emphasize that every research project must define its "communicative purposes". We argue that lexical and syntactic features alone may not suffice for this purpose.

Swales (1990) and, a little later, Bhatia (1993b) were the first to focus on genre analysis based on move structure. Analysis using move structure is a text analysis method developed by Swales in 1981 as an essential component of his genre anal-

LREC-COLING 2024, pages 3916–3926

ysis framework (Swales, 1990). Moves are *"discursive or rhetorical units fulfilling coherent communicative functions in texts"*, whose linguistic realizations can vary widely in length and other ways (Swales, 2004, p. 228-229). They are associated with the notion of steps, which are the multiple fragments of text that *"together, or in some combination, achieve the move"* in such a way that *"the steps of a move function primarily to achieve the goal of the move to which they belong"* (Biber, 2007, p. 24). However, a prototypical schematic structure will be recognizable in terms of the most typical pattern of realization, as identified by the discourse community (Swales, 1990, p. 55). One of the most interesting results in this research's field is to explore the *"common repertoire"* of rhetorical strategies, i.e. all the different possibilities that exist for saying practically the same thing (achieving the same move or communicative goal), and to determine whether certain expressions are more preferred, and therefore more genre-specific, than others. The common method of move analysis can be performed with a bottom-up approach (searching for lexical, grammatical or syntactic features to characterize the move) or a top-down approach (close reading of the text for topic breaks or shifts in content (Moreno and Swales, 2018)).

Move analysis must be performed manually (Biber, 2007), as interpreting communicative functions is a cognitive task that is difficult to access and operationalize (Moreno and Swales, 2018). Indeed, this assumption partly explains why most analyses to date tend to focus on small corpora, as manual analysis is time-consuming, resulting in what some consider a gap in move analysis, namely the lack of elaborate quantitative studies (Biber, 2007).

2.2. Automatic Genre Identification

Several studies have been carried out on automatic genre identification using quantitative and statistical techniques. Most of them use Biber's work but implement a wide range of variants (Santini, 2004). In fact, Biber's work is considered as a pioneering work in the field of automatic genre identification. Biber (1988) makes a distinction between genre and text type, genre is based on external, non-linguistic, traditional criteria and text type is based on the internal, linguistic characteristics of texts themselves. He later analyzed 23 spoken and written genres, with a multidimensional analysis and a quantitative method on 67 different lexical and grammatical features. In 1995, he analyzed a corpus of various spoken and written genres in four languages; this time with a multidimensional analysis and cross-linguistic analysis and a quantitative method on numerous lexical and grammatical features. Later, with Conrad (Biber and Conrad, 2019), they studied various historical and contemporary genres (scientific prose, TV series) with the same approach of multidimensional analysis and quantitative method. One of the first research based on Biber's work is as Karlgren and Cutting (1994), who analyzed the Brown Corpus to explore third person pronouns in text, with discriminant analysis.

A lot of studies explored syntactic features (PoS tags, counting, nominalization) or type-token ratios, or simple token-level measures (Gonçalves, 2021). According to Gonçalves (2021), the classification models often are multiple regressions, Naïve Bayes, discriminant analysis and Recurrent Neural Networks (RNNs) (Stamatatos et al., 2000; Dewdney et al., 2001; Lee and Myaeng, 2002). More recently, Schulman and Barbosa (2018) tried to analyze only the PoS features for English text genre classification. Their aim was to answer whether PoS patterns exist within a text that can distinguish it from one of a different genre. The experiments showed that text genre, which is based solely on lexical content, correlates with the syntactic patterns of the language used. The research achieved up to 80% accuracy in differentiating technical reviews and humanist poetry based on the grammatical role of the language they employ.

Not many papers have used move structure in automatic written genre identification. A recent survey (Kuzman and Ljubešić, 2023) discusses extensively the theoretical inspirations for automatic genre identification, and move structure is not one of them. We could find only two natural language processing paper exploring move structure. Wu et al. (2006) use a method for computational analysis of move structures in abstracts of research articles. They experiment with ways of automatically analyzing the move structure of English research articles abstracts, by (1) building a language model of abstracts moves, then (2) present a prototype concordance which exploits the move-tagged abstracts for digital learning. It seems that automatic detection of move structure is not common in this field. There is another attempt of Dayrell et al. (2012) in automatically identifying rhetorical moves in scientific texts. The authors introduce MAZEA (Multi-label Argumentative Zoning for English Abstracts), a multi-classifier to identify moves.

However, theories of moves (Swales, 1990; Bhatia, 1993b) are often applied in English for specific purposes (ESP) or corpus linguistics. We believe that the absence of exploration of this theoretical framework in the natural language processing literature represents a gap.

2.3. Automatic identification of Web Genre

At a time of widespread digitization of society, after written and oral genres, web genre or digital genre needs to be investigated, characterized and systematically described. This section describes studies of automatic identification of web genre.

Stamatatos et al. (2001) study on various Greek Web sites, by using an NLP tool called SCBD, which proposes 22 parameters or style markers, and information from chunking and parsing. They use a set of style markers for analyzing texts with an NLP tool, using three stylometric levels: token-level, phrase-level, and analysis-level.

In the project TyPWEB, Beaudouin et al. (2002) aim at characterizing commercial websites and personal homepages, by looking at textual, structural, and presentational features. They use word-counting, HTML tags, multivariate statistics, and NLP tools to analyze personal pronouns, grammatical words, lexical opposition, and links.

Santini (2004) made a criticism of Biber's work in automatic genre detection. According to her, the genre typology of Biber is confusing, as he doesn't give a clear distinction between "text type" and "genre"; and the list of 67 features is mostly at the syntactic level. First, according to Santini, who was inspired by the Swalesian definition of genre, the main goal of genre identification is to identify groups of texts that share a common form of transmission, purpose, and discourse properties (Santini, 2004). Automatic genre categorization, in her opinion, "is based on a quantitative approach, leveraging on extractable and computable features (i.e. observable properties in a text) to discriminate among different classes of documents". In addition, Santini (2006) presents three typical features of a genre: hybridism, individualization, and evolution. Genres, according to her, "are not mutually exclusive and different genres can be merged into a single document, generating hybrid forms". Also, because genre can allow a certain freedom of variation and can be dynamic and change over time, Santini (2006) redefines genre of written texts as "named communication artefacts characterized by conventions, raising expectations, showing hybridism and individualization, and undergoing evolution. This definition matches well with written genre in web document, as Santini (2007) proposes a characterization of genre for automatic genre identification of Web pages. She presents an inferential model based on a modified version of Bayes' theorem called odds-likelihood or subjective Bayesian method. The model aims to capture genre hybridism and individualization in web pages, and it uses a combination of linguistic features, HTML tags, and text types as attributes to define genres. The research also describes the steps involved in the model, such as representation of the web in a corpus, extraction and normalization of genre-revealing features, and calculation of probabilities and weighted features.

3. Resource Creation

3.1. Data

For the purpose of our work, we collect a corpus of 120 Tourism website homepages, in French. Indeed, this work is part of a larger project focused on tourism websites in French. This corpus includes 4 genre sub-categories: 30 travel agencies (TA) websites, 30 Travel blogs (TB), 30 websites of Public Tourism Information (PTI) and 30 websites that present Points of Interest (POI). The selection of websites for each sub-category was randomly performed out of a larger corpus made for the broader project this work is a part of. All corpora contain websites made by entities or persons from Belgium, Canada, France, and Switzerland.

It is necessary to underline both the similarity and the distinctiveness of these 4 sub-corpora. First of all, in terms of content, the broad themes addressed in these four corpora are relatively the same (presentation of the destination(s), tourism events, etc.), as they all address tourism. However, each sub-corpus has its own characteristics, especially in terms of their communication purpose. For example, TA conveys promotional/commercial discourse, the sender being private-sector travel agencies, the purpose of which is to sell tourism products/services. PTI is more concerned with institutional discourse, providing information about specific geographical locations (e.g. the Wallonia region in Belgium or the city of Montreal in Canada), with the aim of making the place attractive. POI is quite similar, with more precise locations (such as museums or castles for example). TB, for its part, displays purely personal discourse and combines two goals of the TA and PTI: providing information and/or advertise services.

3.2. Annotation Guidelines

In this section, we give an overview of how we described the annotation process to the annotators.

First, in order to annote the corpora, we present a series of steps to do in our annotation guidelines:

- Step 1: Open the website's link on your desktop PC browser, using full screen mode.
- Step 2: Identify each block of text on the home page, following the reading order (left to right, top to bottom). To identify blocks of text on

a home page, we adopt the following two approaches: (1) based on the web page's visuals/interface, analyze the visual layout of the web page and identify areas that contain text; (2) verify the consistency of those areas by reading each one, analyze the semantic content to identify text blocks according to their meaning or context, to ensure textual cohesion within the text block.

- Step 3: Identify the move type for each distinct text block.

The annotation instructions for move identification are based on the move structure terminology proposed by Askehave and Nielsen (2005), which is dedicated to web pages analysis. However, we have changed and adapted the definitions of few moves, as the definition proposed by the authors is sometimes too vague for our purpose. In addition to the definition, we include accompanying information for each move, such as examples and a brief flowchart, to help annotators understand the move definition. Here is the description of each move:

- Move 1: **Attracting attention**. The aim of this move is to attract the reader's attention as they enter the home page.

- Move 2: **Greeting**. The purpose of this move is to welcome Internet users. It accentuates the metaphor of the home page door: it is set to give the impression of greeting someone on the doorstep.

- Move 3: **Identifying sender**. This move serves to identify the web-owner and is often achieved through a logo. Even though the logo most often contains a purely visual element along text (slogan or name), we only annotate the textual part of the unit.

- Move 4: **Indicating content structure**. It provides the web user with a clear overview of the content of the web site. It is often referred to as the main menu. In our annotation process, we have chosen to annotate all menus displayed on the home page, since every type of menu has the function of indicating the content structure on the site.

- Move 5: **Detailing (selected) content**. This move offers more detailed information about the topics listed in the main menu. It represents the main informational content of the web page. Apart from detailing information, this move also functions as a device for news presentation and public image creation, as news of various kinds seem to be the preferred content of this move (be it international/national news or news of the self-promotional kind, such as financial results, product news, or latest events where the web-owner is involved). It can also be a presentation of detailed information about the company/organization/person: professional activities, services offered, etc.

- Move 6: **Establishing credentials**. This move is meant to establish a trustworthy image of the web-owner. For this move, we had to provide more details to guide annotators. We then based on rhetoric analysis, with Aristotle's theory of *ethos*, *logos* and *pathos*. This theory has been applied into a lot of persuasive discourse, which is logical in the case of the actual move. Thus, this move can be defined by (1) rational, logical discourse supported by figures, graphs, or percentages (logos), (2) speeches promoting the brand/product/services and its reputation (ethos), and (3) speeches highlighting the advantages for Internet users/consumers of coming to the site or taking advantage of the services offered on the site (pathos).

- Move 7: **Establishing contact**. This move encompasses ways for the reader to contact the sender.

- Move 8: **Establishing a (discourse) community**. This move enables loyal or frequent web users to establish communities revolving around the web site (often realised through a private user space or interaction through external social networks).

- Move 9: **Promoting an external organization**. This move promotes another company, product, etc. It usually takes the form of a banner advertisement.

3.3. Annotation Process

Two annotators – among the authors of this paper – participated in the manual annotation of moves. First, as a training step, they annotated separately the same five website homepages, then met (with yet another author of this paper) to (1) get a first idea of the agreement between them, (2) discuss the guide's limitations and (3) discuss enhancements to the guide for the next annotation phase. The same step was performed with another five website homepages, in order to evaluate the agreement and finalize the annotation guide. Once the annotation guide was established, annotators worked on twenty other home pages in the corpus, then we calculated the inter-annotator rate. The inter-annotator agreement rate – Cohen's κ

(Cohen, 1960) – is 0.73, suggesting that the guidelines can yield a reliable annotation. Note that for cases of disagreement on move boundaries, we counted as many items as produced by the annotator who split the most, and entered 0 as a category for each extra item. For example if annotator A identifies one block as being a move 5 where annotator B identifies three such separate items, there will be one agreement and two disagreements. When only focusing on the labels and not penalizing segmentation disagreements, the inter-rater agreement goes up to 0.82.

3.4. Resulting Resource

In this section, we describe the result of the annotation of the corpus. Table 1 shows an overview the size of the 120 documents that were annotated, and the length of the moves that were identified in the process, in tokens.

The means are mostly in the same range for all move types. The standard deviations are very high, most often greater than the means. Those observations indicate that there is no consistency between the text genre and the lengths of the moves. Regarding the functional level of discourse, we can observe that the same moves are important in all corpora (e.g. moves 4 and 5 represent the most and move 2 the least). This suggests that surface observations are not a good indication of the boundaries between the genres that are contained in our corpus.

4. Experiments

With the annotated corpus, we perform experiments on the task of automatic genre identification. Our goal is to test the following hypothesis: informing a model with move structure can increase a model's performance for this task.

We first present the experimental protocols that we put in place (Section 4.1) before presenting the results (Section 4.2). We will conclude the section with a discussion (Section 5).

4.1. Experimental Protocol

We perform three different experiments that we introduce here.

The first one is our baseline: we use the French language model CamemBERT (Martin et al., 2020) to get representations for our documents – for one document, an average of all the corresponding word piece embeddings, excluding paddings, of the last layer produced by the model – and use those as an input for a multi-layer perceptron (MLP). We choose this representation technique for our baseline as our intention is to apply a well-know standard technique for embedding the moves. This way of doing so has been shown to perform consistently better than using the CLS token for various types of sequences (Reimers and Gurevych, 2019; Huang et al., 2021). It has also been done in the task of automatic genre identification (Dömötör et al., 2022). Regarding the model's hyperparameters, we choose to set the learning rate at 1e-05 and to not use dropout, after a grid search on the baseline task. The ranges for the grid search go from 0 to 0.9 by steps of .1 for a dropout layer on the input and before the output, separately, and from 1e-07 to 1e-03 for the learning rate. The hyperparameters are the same for all experiments. We choose to train the hyperparameters on the baseline as a way of reinforcing it, as it is a really standard and simple approach. All layers of CamemBERT are frozen for all the experiments.

The second protocol is the main experiment for testing our hypothesis – which we subsequently refer to as `MOVE`. The only difference lies in how the documents are represented. We replace the single representation of the documents with what follows. During input data creation (i.e. encoding through CamemBERT tokenizer), we group all move segments, belonging to one given document, by move category (i.e. all texts labeled move 1 are concatenated, all texts labeled move 2 are concatenated, and so on). We produce one representation for each move set obtained this way – an average of all the word piece representations for a given combination of document and move –, hence 9 different embeddings for one document. We concatenate those embeddings in order to obtain a single representation for the document (i.e. nine vectors of dimension 768 concatenated into a vector of dimension 6,912). We feed those representations to the same MLP architecture as the one used in the baseline (adjusting for the input size).

This approach for representation enables us to feed more text into the model than the baseline, due to the language model's limitation of 512 word pieces maximum for the input. Indeed, as seen in Table 1, our documents may often exceed this limit (as word pieces are more numerous than tokens). In consequence, we also run experiments to test whether the discrepancy we observe between the two protocols is actually due to the information on move structure, or simply to the increase in text volume per document that can be fed to the model. In order to check for this, we run the same experiment but we randomize the order in which the nine moves are concatenated, document by document. This means that while in our main protocol the first 768 dimensions of the representation of a document are always representations of the concatenation of all moves labeled 1, the next 768 always

Text	TA	TB	PTI	POI
Doc.	419.97 ± 273.18	659.57 ± 478.97	340.61 ± 211.80	277.06 ± 183.20
Move 1	13.86 ± 11.85	11.61 ± 7.65	10.83 ± 12.31	14.90 ± 12.31
Move 2	5 ± 0	61.83 ± 43.18	80.40 ± 114.75	92.29 ± 73.73
Move 3	10.67 ± 26.94	12.35 ± 18.26	7.35 ± 6.24	7.84 ± 7.40
Move 4	7.60 ± 8.26	9.42 ± 11.30	10.61 ± 11.90	7.99 ± 7.99
Move 5	43.81 ± 46.82	48.71 ± 44.46	26.63 ± 38.10	38.32 ± 58.75
Move 6	37.93 ± 28.35	52.40 ± 51.60	31.43 ± 48.74	19.70 ± 19.52
Move 7	18.90 ± 26.21	15.20 ± 17.93	14.17 ± 16.19	17.60 ± 20.09
Move 8	7.95 ± 7.48	17.15 ± 18.65	9.11 ± 9.50	13.31 ± 27.61
Move 9	12.05 ± 10.68	17.24 ± 18.37	30.78 ± 28.37	11.82 ± 6.63

Table 1: Average length, with standard deviation, by move category, in tokens. Column headers are genres: TA = Travel Agencies, TB = Travel Blogs, PTI = Public Tourism Information, POI = Points of Interest

represent the concatenation of all moves labeled 2, and so on, here the order of the move labels randomly changes for each document. We refer to this method as `MOVE Shuffle`.

Figure 1 summarizes the three main models that are used for our experiments. For each experiment, we use 80% of the corpus for training and 20% for testing, randomly with stratification.

As a last note, it may be the case that CamemBERT received part of our data, if not all, during its training. We believe it does not impact our conclusions, as our goal is to compare results between approaches in order to study what move structure can bring to a neural architecture, and not to propose the best performing approach for the task.

4.2. Results

In this section, we introduce the results for the three sets of experiments. Each reported result is the average over 10 iterations. Results of the three experiments are available in Figure 2. We can see that the baseline is able to learn the task and eventually reach high accuracy. As far as our hypothesis is concerned, it is clear that incorporating the move structure into the model reduces the number of epochs required to reach the same degree of performance, even slightly above: it takes the baseline more than 600 epochs to reach `MOVE`'s performance after 100 epochs. The results regarding the control for text volume (`MOVE Shuffle`) show that the move structure does have an effect, as having move types unaligned – due to randomization – clearly decreases the performance of the model, which plateaus at slightly above .70 where the two other methods go beyond .90.

The faster learning capacity of the `MOVE` approach led us to check whether it would reach high performance with a very small sample of data. To that end, we ran the baseline and the MOVE approach with 25% for train and 75% for testing, instead of the 80/20 that we used for the previous set of experiments. The results can be seen in Figure 3. It clearly appears that there is a slight reduction of performance (average of .84), the baseline never reaches `MOVE`'s accuracy (it plateaus at .80). In consequence we can observe that the proposed approach reduces the model's data hunger, being able to greatly reduce the numbers of epochs with as few as 30 documents for training the MLP on top of CamemBERT embeddings.

5. Future Work

Our experiments enabled us to verify our hypothesis: move structure has a noticeable positive impact on the performances of the model within the context of genre identification. Our third set of experiments allows us to rule out the quantity of text fed to the model as an explanation for why the performance increase. This indicates that automatically segmenting and labeling move structure within documents is a promising research venue for automatic genre identification.

6. Discussion

6.1. Genre Complexity

If genre studies from all disciplines share one thing in common, it's the complexity of genres. Whether we choose to analyze genres in terms of textual features, social actions, communities of practice, power structures or the networks and modalities in which they operate (and individual researchers must almost always limit themselves to certain dimensions among these), we know that we have only a partial view of all that is really going on (Johns et al., 2006). So, whichever approach we follow, the ways in which we analyze genre are only partial representations of the complex nature of genre and the social and communicative functions they have to fulfill.

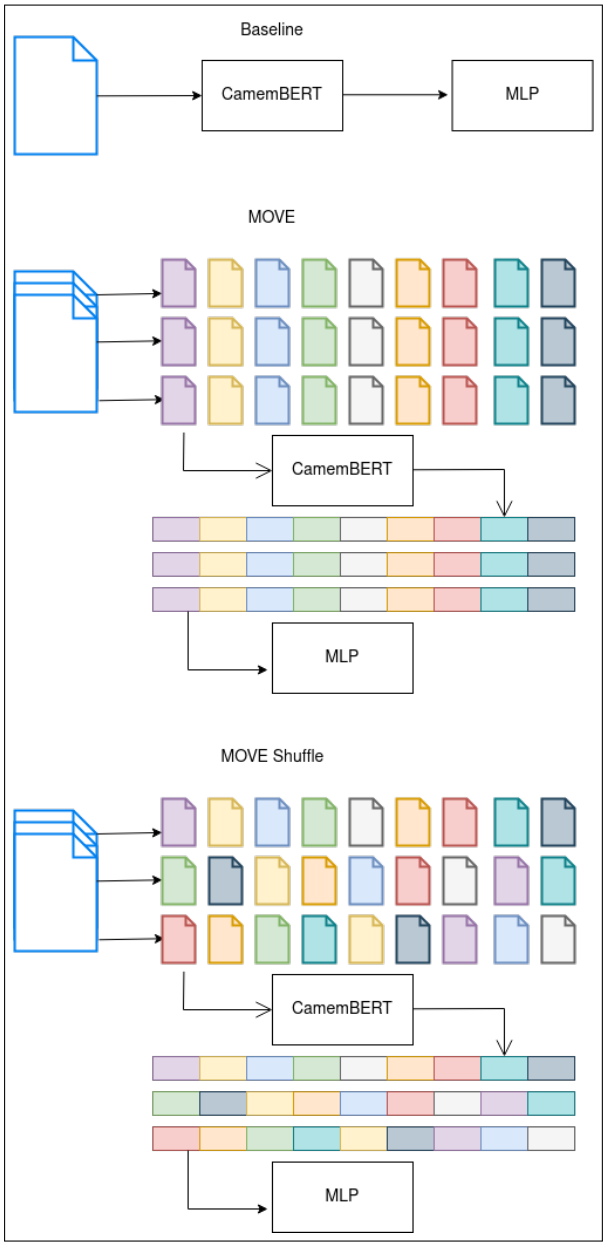

Figure 1: Summary of the three models used in our experiments. Each color denotes a distinct move type.

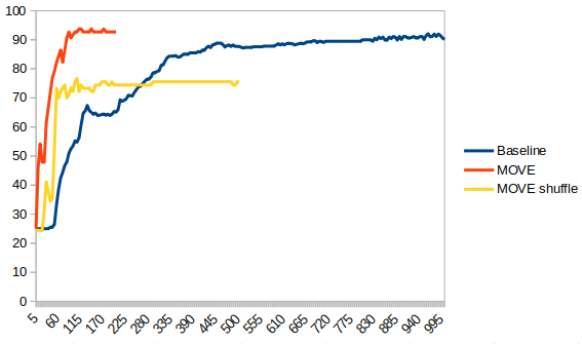

Figure 2: Results for the three experiments, Baseline (CamemBERT + MLP), MOVE (CamemBERT with move information + MLP) and MOVE shuffle (CamemBERT with shuffled move information + MLP). X-axis is the number of epochs, Y-axis is test accuracy.

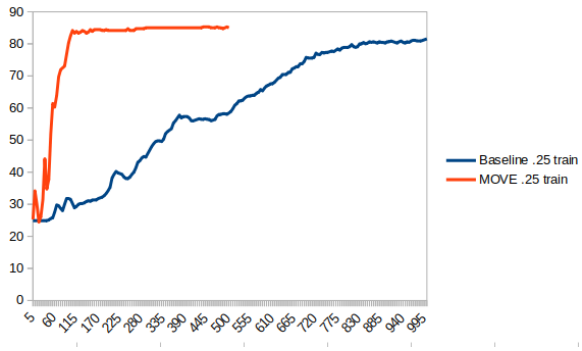

Figure 3: Results for Baseline (CamemBERT + MLP) and MOVE (CamemBERT with move information + MLP) with 25% for training and 75% for testing. X-axis is the number of epochs, Y-axis is test accuracy.

In digital genre analysis, despite the different points frameworks focus on, they address a common theme of digital genres, namely the increased interconnectedness between discourses. While it is common for genre analysts to place one genre in relation to other genres, this practice has become more complicated due to digital technologies that allow a text to be easily linked to other texts through hyperlinks. This function of hypertextuality provokes possible ways of constructing genres, and at the same time allows readers to consume genres in their own personalized ways. This possibility therefore obliges analysts to examine the differences created by hyperlinked content in achieving the genre's communicative goals, particularly in comparison with a non-digital genre without hyperlinks. Another problem is the new, complex type of digital discourse – tourism discourse in this case in particular – we need to revise or revisit the so-called traditional genre models (Askehave and Nielsen, 2005). Moreover, when analyzing digital discourse in the field of tourism, we need to take a multidimensional approach, paying attention not only to the text, but also to other elements that are linked to the text (hyperlinks, visual graphic elements, but also other non-textual yet discursive and commercial elements).

6.2. Web Genre Typology

A possible criticism of Askehave and Nielsen's framework is that it does not provide a clear explanation of how the concept of move might be operationalized in a text containing hyperlinks, which could interrupt a traditionally defined rhetorical unit

of move (Mehlenbacher, 2017). To address this criticism, analysts may need to consider that digital genres offer readers different reading paths (Baldry et al., 2006) and that a reader may choose to read according to a dominant mode used in a text (Kress, 2003). For example, when reading a home page containing a large part of the text as well as a hypertext link, a likely reading path would be to read the full text first, before clicking on the hypertext link, which would lead to a different rhetorical structure.

6.3. Annotation Process

During the annotation process, we encountered two main problems: the establishment of the annotation guide based on studies by Askehave and Nielsen (2005) and the manual annotation itself. Regarding the first obstacle, the proposed typology of Askehave and Nielsen (2005) leaves room for interpretation, it is not detailed enough to be able to use it as annotation guidelines as it is. Indeed, the authors emphasize in their article that this is an attempt to (1) apply Swales' model of move structure to the analysis of websites (home pages in particular), (2) propose an analytical framework for genre analysis that doesn't reduce media-specific elements to something beyond the genre itself. Their aim is not to establish a taxonomy of digital genres and/or characteristics specific to digital text genres, their analysis is purely interpretative and based on their own observations as default readers of websites. In consequence, we had to adapt our annotation guidelines, by expanding or reducing some categories of their typology (e.g. adding marketing content to move 6).

Two obstacles were revealed during the manual annotation: the text segmentation and a series of "borderline cases" where the segmented text can belong to two moves, which point to a certain degree of subjectivity. First, it has become common practice to segment texts into moves, these being considered as discourse units containing at least one proposition (e.g. Connor and Mauranen (1999)). However, some studies have equated this notion with grammatical units such as a sentences or paragraphs (e.g. Hopkins and Dudley-Evans (1988); Peacock (2002)). We can see that approaches for text segmentation are various, and strongly depend on the research objectives of each researcher. We had then to decide how annotators could segment texts in the home page interface and came up with the idea of identifying "block of text", based on visual analysis.

For the second obstacle, few ambiguous cases have been identified, despite the efforts made to clarify the annotation guide. Below is an example that annotators labeled as move 5 or move 6:

Figure 4: Example of text segmentation's ambiguity. – English translation: *Support the blog! If you'd like to help me keep this travel blog going, please feel free to make purchases and bookings via the links and advertising blocks. I earn a small commission to keep this blog going, and the prices are exactly the same for you. Thank you in advance! Book a hotel Buy a plane ticket Rent a car Buy on Amazon Buy a train ticket)*

"Resorts exclusifs, adresses confidentielles, city trips inspirants, croisières de luxe ou circuits exotiques vers les destinations les plus extraordinaires du monde – laissez-vous inspirer par le portefeuille de voyages premium de SIGNATURE VOYAGES." (English translation: "Exclusive resorts, confidential addresses, inspiring city trips, luxury cruises or exotic tours to the world's most extraordinary destinations - let yourself be inspired by SIGNATURE VOYAGES's premium travel portfolio") Indeed, while this is a text that provides detailed information about the services offered by the company (move 5), it also includes promotional terms about the brand, its reputation, and the advantages for Internet users of taking advantage of the services offered (move6). Regarding the incorporation of move structure in natural language processing tasks, this may suggest that it would be beneficial to take those borderline cases during annotation in a way or another (e.g. allowing multi-label or creating specific labels for those borderline cases).

Another example is about text segmentation ambiguity, due to the website owner's decision on information presentation, which can be seen in Figure 4.

We can see that the information is presented in the same block. However, in terms of textual content annotators could not decide if this is either

move 4 (indicating a structure) or move 5 (detailing selected content).

7. Conclusion

In this paper, we conducted a study on the task of automatic genre identification, working specifically with websites related to tourism, in French. We introduced a new corpus annotated with move structure, following an existing typology that we confronted to data we collected online. We showed that our corpus could be used for experiments on automatic genre identification. We could show that leveraging move structure can help models learn better and faster, while reducing data hunger. Our insights make the case for investigating the task of automatic move identification, which is a line of research that has received very little attention from the community.

8. Acknowledgements

We would like to thank the anonymous reviewers for their questions and comments that helped improve the quality of this paper. Computational resources for the experiments described in this paper were provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

9. Bibliographical References

Jean-Michel Adam. 1997. Genres, textes, discours: pour une reconception linguistique du concept de genre. *Revue belge de philologie et d'histoire*, 75(3):665–681.

Jean-Michel Adam. 1999. Linguistique textuelle: des genres de discours aux textes. *(No Title)*.

Inger Askehave and Anne Ellerup Nielsen. 2005. Digital genres: a challenge to traditional genre theory. *Information technology & people*, 18(2):120–141.

Anthony Peter Baldry, Paul J Thibault, et al. 2006. *Multimodal transcription and text analysis: A multimedia toolkit and coursebook*, volume 1. Equinox.

Jean-Claude Beacco. 2013. L'approche par genres discursifs dans l'enseignement du français langue étrangère et langue de scolarisation. *Pratiques. Linguistique, littérature, didactique*, (157-158):189–200.

Valérie Beaudouin, Serge Fleury, Benoît Habert, Gabriel Illouz, Christian Licoppe, and Marie Pasquier. 2002. Typweb : décrire la toile pour mieux comprendre les parcours. *Réseaux*, 116.

Vijay K Bhatia. 1993a. Language use in professional settings. *Applied Linguistics and Language Study.) London: Longman*.

V.K. Bhatia. 1993b. *Analysing Genre: Language Use in Professional Settings*. Applied linguistics and language study. Longman.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.

Douglas Biber. 2007. *Discourse on the move: Using corpus analysis to describe discourse structure*, volume 28. John Benjamins Publishing.

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press.

Jean-Paul Bronckart and Joaquim Dolz. 2002. La notion de compétence: quelle pertinence pour l'étude de l'apprentissage des actions langagières. *Raisons éducatives*, 2:27–44.

Patrick Charaudeau. 2011. Chapitre 1. l'information comme acte de communication. *Medias-Recherches*, 2:21–28.

SG Chartrand, J Émery-Bruneau, and K Sénéchal. 2015. Caractéristiques de 50 genres pour développer les compétences langagières en français au secondaire québécois didactica, céf. *Québec: Université Laval*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Ulla Connor and Anna Mauranen. 1999. Linguistic analysis of grant proposals: European union research grants. *English for specific purposes*, 18(1):47–62.

Carmen Dayrell, Arnaldo Candido Jr., Gabriel Lima, Danilo Machado Jr., Ann Copestake, Valéria Feltrim, Stella Tagnin, and Sandra Aluisio. 2012. Rhetorical move detection in English abstracts: Multi-label sentence classifiers and their annotated corpora. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1604–1609, Istanbul, Turkey. European Language Resources Association (ELRA).

Nigel Dewdney, Carol VanEss-Dykema, and Richard MacMillan. 2001. The form is the substance: Classification of genres in text. In *Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management*.

Andrea Dömötör, Tibor Kákonyi, and Zijian Győző Yang. 2022. What's your style? automatic genre identification with neural network. *Computación y Sistemas*, 26(3):1293–1299.

A Gonçalves. 2021. A supervised text mining approach for automatic text genre classification. In *16th Doctoral Symposium in Informatics Engineering*, page 168.

Andy Hopkins and Tony Dudley-Evans. 1988. A genre-based investigation of the discussion sections in articles and dissertations. *English for specific purposes*, 7(2):113–121.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. WhiteningBERT: An easy unsupervised sentence embedding approach. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 238–244, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ann M Johns, Anis Bawarshi, Richard M Coe, Ken Hyland, Brian Paltridge, Mary Jo Reiff, and Christine Tardy. 2006. Crossing the boundaries of genre studies: Commentaries by experts. *Journal of second language writing*, 15(3):234–249.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Gunther R Kress. 2003. *Literacy in the new media age*. Psychology Press.

Taja Kuzman and Nikola Ljubešić. 2023. Automatic genre identification: a survey. *Language Resources and Evaluation*, pages 1–34.

Yong-Bae Lee and Sung Hyon Myaeng. 2002. Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–150.

Dominique Maingueneau. 2004a. *Le discours littéraire: paratopie et scène d'énonciation*. Armand Colin.

Dominique Maingueneau. 2004b. Retour sur une catégorie: le genre. *Texte et discours: catégories pour l'analyse*, pages 107–118.

Dominique Maingueneau. 2007. Genres de discours et modes de généricité. *Le français aujourd'hui*, (4):29–35.

Dominique Maingueneau. 2016. *Les termes clés de l'analyse du discours*. Média Diffusion.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Ashley Rose Mehlenbacher. 2017. Crowdfunding science: Exigencies and strategies in an emerging genre of science communication. *Technical Communication Quarterly*, 26(2):127–144.

Ana I. Moreno and J. M. Swales. 2018. Strengthening move analysis methodology towards bridging the function-form gap. *English for Specific Purposes*, 50:40–63.

Matthew Peacock. 2002. Communicative moves in the discussion section of research articles. *System*, 30(4):479–497.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Jean-Jacques Richer. 2011. Les genres de discours: une autre approche possible de la sélection de contenus grammaticaux pour l'enseignement/apprentissage du fle? *Linx. Revue des linguistes de l'université Paris X Nanterre*, (64-65):15–26.

Marina Santini. 2004. State-of-the-art on automatic genre identification. Technical report, (Technical Report ITRI-04-03). Information Technology Research Institute.

Marina Santini. 2006. Some issues in automatic genre classification of web pages.

Marina Santini. 2007. Automatic genre identification: Towards a flexible classification scheme.

Alan Schulman and Salvador Barbosa. 2018. Text genre classification using only parts of speech. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1226–1229.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 2000. Text genre detection using common word

frequencies. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

Efstathios Stamatatos, Nikos Fakotakis, and Kokkinakis George. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35:193–214.

J.M. Swales. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge Applied Linguistics. Cambridge University Press.

John M Swales. 2004. *Research genres: Explorations and applications*. Cambridge University Press.

Amalia Todirascu. 2019. Genre et classification automatique en tal : le cas de genres journalistiquesgenre and nlp : the case of the automatic classification. *Linx*.

Jien-Chen Wu, Yu-Chia Chang, Hsien-Chin Liou, and Jason S. Chang. 2006. Computational analysis of move structures in academic abstracts. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 41–44, Sydney, Australia. Association for Computational Linguistics.