2012-03-05

A Comparative Analysis of Text Usage and Composition in Goscinny's *Le petit Nicolas*, Goscinny's *Astérix*, and Albert Uderzo's *Astérix*

Dennis Scott Meyer
Brigham Young University - Provo

A Comparative Analysis of Text Usage and Composition in

Goscinny's *Le petit Nicolas,* Goscinny's *Astérix*,

and Albert Uderzo's *Astérix*

Dennis Meyer

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Yvon LeBras, Chair
Deryle Lonsdale
Nicolaas Unlandt

Department of French and Italian

Brigham Young University

April 2012

ABSTRACT

A Comparative Analysis of Text Usage and Composition in
Goscinny's *Le petit Nicolas*, Goscinny's *Astérix,*
and Uderzo's *Astérix*

Dennis Meyer
Department of French and Italian, BYU
Master of Arts

The goal of this thesis is to analyze the textual composition of René Goscinny's *Astérix* and *Le petit Nicolas*, demonstrating how they differ and why. Taking a statistical look at the comparative qualities of each series of works, the structural differences and similarities in language use in these two series and their respective media are highlighted and compared. Though one might expect more complicated language use in traditional text by virtue of its format, analysis of average word length, average sentence length, lexical diversity, the prevalence of specific forms (the *passé composé*, possessive pronouns, etc.), and preferred collocations (*ils sont fous, ces romains !*) shows interesting results. Though *Le petit Nicolas* has longer sentences and more relative pronouns (and hence more clauses per sentence on average), *Astérix* has longer words and more lexical diversity. A similar comparison of the albums of *Astérix* written by Goscinny to those of Uderzo, paying additional attention to the structural elements of each album (usage of narration and sound effects, for example) shows that Goscinny's love of reusing phrases is far greater than Uderzo's, and that the two have very different ideas of timing as expressed in narration boxes.

Keywords: René Goscinny, Albert Uderzo, Astérix, Le petit Nicolas, lexical diversity, TreeTagger, part of speech tagging, lemmatization, average word length, average sentence length, verb tense choice, preferred collocations, comics, les bandes dessinées

ACKNOWLEDGEMENTS

For their encouragement and support, I thank Jackie Meyer, Adam McBride, and Jay Carr. For their assistance in corpus construction, I thank Donna Andrus, Jeff Asay, Haleigh Burgon, Kirby Childress, Andrew Current, Bethany Daniel, Annie Fisher, Adam McBride, Susan Garver, Spencer Greenhalgh, Amy Hansen, Kevin Holt, Irene Kim, Nathan Oaks, and Ben Sparks. For their assistance in verification of part of speech tagging, I thank Emily Coyne, Élise Danguy, Jourdan Hafen, Irene Kim, Johnna Kretchman, Adam McBride, Katie Meyer, Joshua Mitton, and Jennifer Trumbo. I would also like to thank professors Laurence Anthony and Yasu Imao for the creation and free distribution of AntConc and CasualTreeTagger respectively.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

The popularity of the *bande dessinée* (or *BD*) in France and Belgium cannot be overstated. So great is the Franco-Belgian love for the form that it has been dubbed *le neuvième art,* joining the ranks of architecture, sculpture, painting, dance, music, and poetry (classifications originally established by Hegel in his *Aesthetics*), as well as fellow newcomers cinema and photography. There, it is a legitimate and celebrated art form. In France, the *BD* industry brought in over 320,000,000 euros in sales in 2007 alone. And that's just the volumes that were sold new. That same year, 1 out of every 20 new books sold in France was a *bande dessinée* (Beuve-Méry). And yet, though French films are nominated for awards the globe over and French literature has been lauded for centuries, the *BD* is, to many international onlookers, seen as no more than a bunch of comic books for children.

In recent years, the *BD* and its cousin the graphic novel have been the subject of many critical studies. Books such as *Lire la bande dessinée* by Benoît Peeters and *Understanding Comics: The Invisible Art* by Scott McCloud are a testament to this on both sides of the Atlantic. Academic journals such as *ImageTexT* have been created, which as the title indicates, have the sole goal of discussing the rich marriage of image and text. However, despite the growing attention that is given to this art form, it seems that the bulk of that attention is paid only to visual and thematic elements, and seldom if ever purely textual ones. Entire books are dedicated to the vocabulary used to describe the graphic layout of a page, articles are written about "visual rhetoric", and elaborate comparisons are drawn between graphic novels and their motion picture adaptations. At a wider scope, the historical and cultural implications of the works are discussed and conferences are held to discuss the portrayal of gender, family and society. Indeed, particular attention is paid to the *genres* of these works, to their post-modern re-mixes of reality and to the overall narrative structure of these works. The list could go on as others discuss the pedagogical merits of such works in teaching grammar, vocabulary, etc. but the point is clear; there are no discussions of *BD* at a purely textual level.

In decades past, many concerned citizens feared that comics could be detrimental to children, as many concerned parents asked themselves questions such as, "Does the language, printing and illustrating (of a comic) impair your child's reading and language skills? Does it foster it?" (Ramage) Many have since claimed that the right comics can indeed foster a child's reading ability, but there is a persistent feeling that remains that can be summed up in a question that the Harvey and Eisner award-winning Jeff Smith relates in the biographical documentary *The Cartoonist*. His mother once asked him, "why don't you read a REAL book?"(Mills) The underlying assumption seems to be this: because comics have pictures, the text MUST suffer as a result. Either the text is qualitatively different in a comic, or else the presence of images somehow demeans it. But is that assessment justified? *Bandes dessinées*, like comic books, are immensely popular with younger readers, and though they certainly contain fewer words *per page* than traditional books, is the *complexity* of the language used comparatively simpler? Is its lexicon poorer? In short, how does the language of the *BD* compare to more traditional texts? What traits define the text used in *BD*?

Certain *bandes dessinées* in particular seem to remain very dear to the hearts of the French and the Belgians. *Astérix* has known an incomparable domestic success in France; only 6 years after the publication of its first volume, the very first French satellite, a source of immense national pride, was given the name "Astérix" (asterix.com "Astérix Galerie - Les Expositions - Les Archives D'albert Uderzo (Suite) - Le Site Officiel"). Older volumes are still being printed, and the series is currently undergoing an updating process to bring the quality of the art and the text up to modern production standards (asterix.com "Astérix Edition - La Grande Collection - Les Étapes De La Restauration - Le Site Officiel"). The series has also known a great international success, having been translated into over 100 languages and dialects (asterix.com "Astérix Encyclopédie - Les Traductions - Le Site Officiel"), proving its universal appeal despite rather obvious Franco-centric nationalistic overtones. Perhaps only *Tintin* has reached a similar apogee of global recognition. But unlike *Tintin*'s author Hergé, who only ever produced *BD*, Goscinny is also known in France for his short stories centering on a young schoolboy known to readers as *le petit Nicolas*.

This presents us with an excellent point of comparison; the fact that both series share a common author means that any variances that could be attributed to a difference in education or style between two authors are minimized. Though the differences in semantic field between a contemporary elementary school boy and Gaulish warriors from 2,000 years ago are likely to be quite large, these differences should be tempered to some extent by the fact that both works were published in the same venue, the children's magazine *Pilote*, with the same audience in mind. Both series continue to sell copies in dozens of languages, and Goscinny was even awarded *le Prix Alphonse Allais* for humor in 1964 for his collection *Le petit Nicolas et les copains*. (Le bulletin du livre, 1964:25) There is, then, perhaps no better point of comparison to see if the medium of the *bande dessinée* itself inevitably leads to a diminished use of the French language.

Another subject of interest presents itself, hiding in the shadows of Goscinny's prose. Since Goscinny's death in 1977, *Astérix* is now written by his long-time collaborator, Albert Uderzo, and though the most recent volumes continue to hit the best-seller charts reported by IPSOS (BD "Top 50 Des Meilleures Ventes Bd En 2006"; BD "Top 50 Des Meilleures Ventes

Bd 2008"), prominent voices in the media claim that Uderzo’s work is comparatively inferior. Anne-Claire Norot, a critic of *bandes dessinées* at *Les Inrocks* magazine, believes that “Goscinny's death was a turning point. After that, the language, the jokes, the subtlety - it was all gone. Before [the Asterix comic] was art, now it is just for children.” (Schofield) Meanwhile, Daniel Schneidermann of the newspaper *Libération* wrote that “le dernier album d’Astérix, hélas, est mauvais” (Schneidermann). Norot and Schneidermann’s specific complaints can be reduced to the feeling that the language is less rich and that the structure of the albums was less coherent. Norot also makes a distinction that rings curiously in many foreign ears: that *Astérix* is now suitable only for children, though *once* it was art, suitable for adults as well.

At a popular level as well, the reception of Uderzo’s work has also been less than kind. At bedetheque.com, where users can rate *BD* themselves, each album of Astérix has been reviewed by at least 59 users, though the most recent was reviewed by over 400. In all of Goscinny’s tenure, his rating never dips below 7.4 out of 10, and this low point (by a .5 margin) was his first album. On average, Goscinny’s albums are rated at 8.6 out of 10. Uderzo’s albums, on the other hand, are rated at an average of 5.2 out of 10, with the lowest receiving only 2.5 out of 10. This low point is the most recent album of which Schneidermann and Norot spoke.

A comparative textual analysis of Goscinny’s *Astérix* and Uderzo’s *Astérix* will shed light on the differences between the two authors’ styles, both in the composition of their text and the composition of their narratives. Though a computer-based analysis will never prove fully satisfactory as a method of comparing the more elegant use of wordplay within a narrative, a statistical analysis of the presence of various elements such as narration (typically contained within square boxes), sound effects (text that occurs outside of dialogue and narration), and the quantity and variety of the text itself should serve to quantify the tendencies of their writing.

2. CORPUS CONSTRUCTION

2.1. Selection of Works to Analyze

2.1.1. Primary Analysis

A nice round target of 100,000 words for each section of the corpus was chosen. Goscinny penned 24 albums of *Astérix* from 1959 to 1977. Initial estimates based on counting the words on random pages throughout the body were that each album weighed in at 7,000 to 8,000 words. Thirteen albums of *Astérix* spanning Goscinny's work were initially selected, with a fourteenth album added once it was discovered that those 13 didn't quite total 100,000 words (see Table 1). A slight predilection towards his earlier work was a conscious choice, given that it would better correspond to the period in which the adventures of *Le petit Nicolas* were written and published.

In contrast to *Astérix*, Goscinny wrote over 200 short stories featuring *Le petit Nicolas* from 1959 to 1965. Using OCR software (further details on this later), each story was estimated at near 1,500 words apiece. As for the selection of which stories to choose from *Le petit Nicolas*, the availability of four specific collections in the university library made the choice quite simple, and their collective 69 short stories totaled just over 107,000 words (see Table 1) in the end.

Albert Uderzo has penned, to date, ten albums of *Astérix*, though this figure is somewhat misleading. Two of those albums took non-standard forms and have hence not been included, the first being dedicated to collecting one-off stories and promotional work over the years (some of which were penned by Goscinny), the second being a sort of retrospective of the *Astérix* franchise that violates the fourth wall, is formatted differently, and is longer than the traditional 44 pages. The eight remaining albums were thus all selected, as their collective weight in words only totals approximately 60,000 words (see Table 1).

Table 1: Works selected for analysis

***Le petit Nicolas* collections written by René Goscinny**			
Number	Title	Year	Words
1	*Le petit Nicolas*	1960	26,645
2	*Les récrés du petit Nicolas*	1961	26,704
3	*Les vacances du petit Nicolas*	1962	31,424
5	*Le petit Nicolas a des ennuis*	1964	22,415
		TOTAL	107,188

***Astérix* albums written by René Goscinny**			
Number	Title	Year	Words
1	*Astérix le Gaulois*	1961	6,067
2	*La serpe d'or*	1962	6,082
3	*Astérix et les goths*	1963	7,421
4	*Astérix gladiateur*	1964	8,285
6	*Astérix et Cléopâtre*	1965	8,070
8	*Astérix chez les Bretons*	1966	8,471
10	*Astérix légionnaire*	1967	7,429
12	*Astérix aux Jeux olympiques*	1968	7,147
14	*Astérix en Hispanie*	1969	7,389
15	*La Zizanie*	1970	7,686
16	*Astérix chez les Helvètes*	1970	7,144
20	*Astérix en Corse*	1973	7,766
23	*Obélix et compagnie*	1976	6,232
24	*Astérix chez les Belges*	1979	7,356
		TOTAL	102,545

Astérix albums by Albert Uderzo			
Number	Title	Year	Words
25	*Le grand fossé*	1980	7,857
26	*L'Odyssée d'Astérix*	1981	8,036
27	*Le fils d'Astérix*	1983	8,155
28	*Astérix chez Rahàzade*	1987	8,038
29	*La rose et le glaive*	1991	7,892
30	*La galère d'Obélix*	1996	7,362
31	*Astérix et Latraviata*	2001	7,304
33	*Le ciel lui tombe sur la tête*	2005	5,963
		TOTAL	60,607

Organized by the sequential number of volumes in each series ("Number"). Final word count figures generated from *Word Counter* by David Hanauer.

2.1.2. Reference Corpus

A small reference corpus was chosen for the purpose of benchmarking the utility of the statistics being generated. Two *BD* volumes, *Tintin au Congo* by Hergé and *Trio de l'Étrange* by Roger Leloup, were chosen by virtue of having the volumes readily available and already typed out (following a reduction in the scale of this project). Two children's books, *Le Petit Prince* by Antoine de Saint-Exupéry and *Le Grand Meaulnes* by Alain-Fournier, were also chosen as reference works against which *Le petit Nicolas* could be measured. These works were chosen for two reasons: they are children's books published within a few decades of *Le petit Nicolas,* and more specifically to reduce preparation time, since both works were accessible in already digitized formats, which saved time preparing the texts via OCR software. The text of *Le grand Meaulnes* was accessed via wikisource.org and the text of *Le petit prince* was accessed via wikilivres.info.

2.2. Text preparation

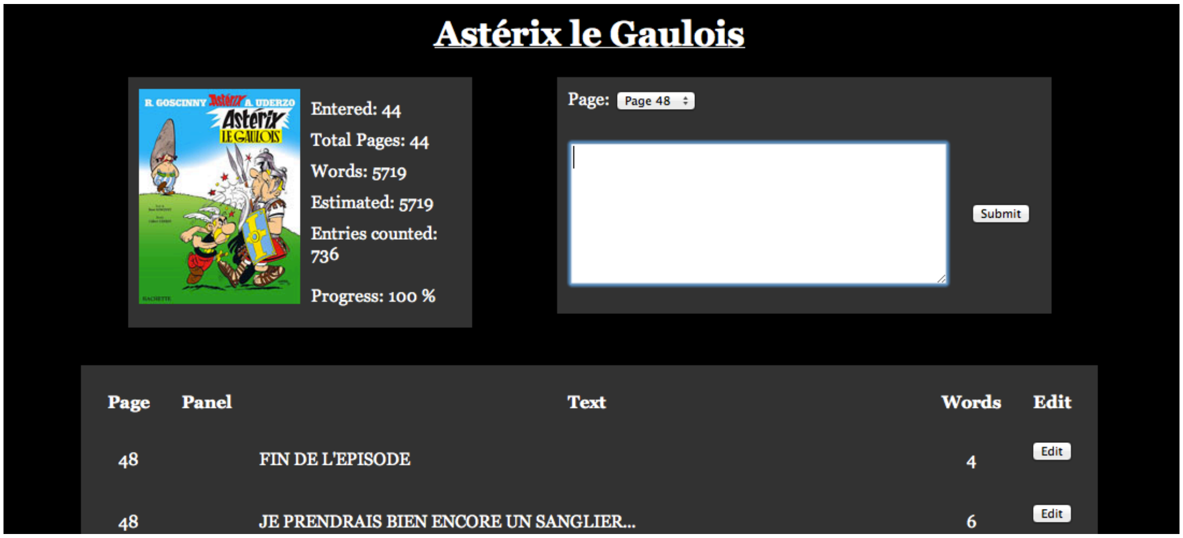
The necessary task of bringing the works into a digital format required a separate method for each series. For the collections of *Le petit Nicolas*, it was sufficient to scan each page and run the resulting images through an OCR (Optical Character Recognition) software solution. *ABBYY FineReader*, a program that was available on the university's library computers, was chosen for practical reasons. It was quite reliable in its reproduction of the works. The only concerns that arose were not textual in nature, but concerned the formatting of the text. Carriage returns were added to the text at the end of each printed line, breaking the flow of sentences. This was only a concern when viewing the texts in such programs as *Microsoft Word*, as the actual analysis of the text ignored such characters to determine sentence length and sequence of words.

Due to limitations in OCR technology, however, any automated entry of text from *Astérix* into a computer is prohibitively impractical. The hand-lettered text of the older comics is too inconsistent for any of the currently available commercial products[1] to assist in the construction of the corpus, and though the newer volumes use standardized machine-set fonts, these also proved unreadable by the software, as the fonts themselves were based on the hand-written characters of the older volumes. The OCR products only produced reliable results from the most common professional fonts, such as the Times New Roman and Arial typefaces. As a result, it was necessary to enter the text of *Astérix* manually, through a combination of speech-to-text software and manual typing. *Dragon Dictate* provided reliable speech-to-text support for the French language to the point that certain sections of the series were much faster to enter by voice than by hand. The software failed in those locations where speech became modified (changes in orthography to reflect a character's accent or drunkenness, for example) and when the frequent use of invented proper nouns punctuated the dialogue.

To aide in the rapid construction of the corpus, I built a MySQL database for the storage of all of the text and any necessary annotations. Storing this database online allowed me the liberty of creating a web interface whereby volunteers (see Acknowledgements) could help enter

[1] *ABBYY FineReader*, *OmniPage Pro*, *Readiris Pro*, and the trainable open source *Tesseract* were all tested. Though the first three did fine at distinguishing when images were embedded in a body of typeset text, none of the products could reliably discern text, even typed text, within the body of a comic. Removing the art digitally to leave only the text produced better results, though the programs were still unable to recognize over half of the hand-written characters from the earliest decades of *Astérix*'s publication.

text. The simple PHP/HTML interface (see Fig. 1) served three purposes: Firstly, it allowed multiple users the ability to enter text simultaneously. Secondly, it generated a preliminary word count for each entry (detecting word boundaries by spaces or punctuation marks), allowing scripts to extrapolate the approximate word count of any given album based on how many words had thus far been entered across how many pages had been completed. This quickly demonstrated that a fourteenth album of *Astérix* written by Goscinny would be necessary. Lastly, it allowed all of the volunteers to have access to the same guidelines for entering text, minimizing any disparities that might otherwise arise by having multiple individuals perform the task.



Astérix le Gaulois

Entered: 44
Total Pages: 44
Words: 5719
Estimated: 5719
Entries counted: 736
Progress: 100 %

Page: Page 48

Page	Panel	Text	Words	Edit
48		FIN DE L'EPISODE	4	Edit
48		JE PRENDRAIS BIEN ENCORE UN SANGLIER...	6	Edit

Figure 1: Custom website for corpus creation and rudimentary word count

Once completed, each *BD* volume's data was exported into a separate XML file (see Fig. 2). These files were then spell checked in *Microsoft Word*, which helped identify much of the human error that occurred in inputting the text, such as typos in the text and entries where the page numbers hadn't been marked properly. Employing *Microsoft Word*'s spellcheck also allowed for a very rapid means of finding the bulk of the entries in need of annotation. Entries were flagged for whether they contained Latin phrases, non-standard text (word deformation to reflect accents, drunkenness, etc.), and onomatopoeias.

```
<?xml version="1.0" encoding="UTF-8"?>
<body>
…
        <entry>
                <id>5</id>
                <page>5</page>
                <text>MAIS ADDENTZION! ON REFIENDRA!</text>
                <gloss>MAIS ATTENTION! ON REVIENDRA!</gloss>
                <sfx></sfx>
                <latin></latin>
                <meanwhile></meanwhile>
        </entry>
…
        <entry>
                <id>1217</id>
                <page>46</page>
                <text>TCHOP!</text>
                <gloss></gloss>
                <sfx>1</sfx>
                <latin></latin>
                <meanwhile></meanwhile>
        </entry>
…
</body>
```

Figure 2: Sample of XML format of annotated text entries

In the XML of figure 2, each entry is broken down into the following tags: the <id> was a unique number for every entry, the <page> tag reflected what page of the album the entry appeared on, the <text> tag showed the exact text in the album, and the <gloss> tag presents standardized French. The <sfx>, <latin> and <meanwhile> tags had only two possible values: 1 or blank. If their values were set to 1, it means that the entry was, respectively, a sound effect (<sfx>), in latin (<latin>), or a narration box (<meanwhile>).

Once edited, the XML files were parsed with a purpose-built PHP script to update the database with any corrections and annotations. The texts of each file were then re-exported in four varieties: Firstly, with the full text of each volume as written. Secondly, as a full text that had replaced any non-standard language usage with standard forms (with such phrases as "*navi'e à t'ibo'd !*" being rendered "*navire à tribord !*"). Thirdly, files were generated with only the sound effects from each volume. Finally, every page had its own text file generated for use in structural analysis of the composition and distribution of language throughout each individual album of *Astérix*.

2.3. Methods of Corpus Analysis

2.3.1. AntConc

With the collected, corrected texts of each work now assembled into text files, two principal tools were used in the analysis of each body of text. *AntConc*, a free program written by professor Laurence Anthony of Waseda University, is a platform-agnostic tool for constructing word lists and various concordance-related statistics. For the purpose of identifying characters in the corpus, only the characters in Figure 3 were recognized as containing word data. This reflects a slightly larger character set than occurs in standard French, but these additional characters were necessary to accommodate Goscinny's propensity for using foreign characters both in sound effects and in creating humorous proper names during Astérix's travels.

abcdefghijklmnopqrstuvwxyzàâåèéëêïîöôøüûæœçñ-

ABCDEFGHIJKLMNOPQRSTUVWXYZÀÂÅÈÉËÊÏÎØÖÔÜÛÆŒÇÑ

Figure 3: Characters recognized as containing word data in *AntConc*

It is important to note that the various works have slightly different attributes. Because *Astérix* is written almost entirely in upper-case letters, while the other works followed normal conventions of capitalization, all tests in AntConc were set to analyze words as though they were entirely in lower-case letters. This minimized any differences that might have arisen from the disparity in capitalization, though the issue of capitalization affected other tests.

To generate more reliable information on the usage of language, a lemma list was employed (UIMA). The UIMA lemmatizer project hosts a massive list of lemmas in French. The text itself was presented in a format that couldn't be used directly in AntConc, presenting a pair of words, the inflected form and its lemmatized form, on each line (see Table 2).

Table 2: Lemma list as formatted by UIMA project

Form	Lemma
laquelle	lequel
lequel	lequel
lesquelles	lequel
lesquels	lequel
leur	leur
leurs	leur

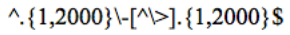
The data was re-formatted into a CSV (comma separated values) file and uploaded into a new table in the MySQL database. From there, a query was run to group all of the inflected forms under shared lemmas. The resulting data was placed into a generated .txt file that fits the model required for AntConc (see Fig. 4).

```
…
étrave->étrave,étraves
être->être,êtres,être,étant,été,suis,es,est,sommes,êtes,sont,étais,étais,était,étions,étiez,étaient,fus,
fus,fut,fûmes,fûtes,furent,serai,seras,sera,serons,serez,seront,serais,serais,serait,serions,seriez,
seraient,sois,sois,soit,soyons,soyez,soient,fusse,fusses,fût,fussions,fussiez,fussent,sois,soyons,
soyez
étréci->étrécie,étrécies,étréci,étrécis
…
```

Figure 4: Sample of formatting for *AntConc* obtained after processing

Because AntConc uses the first listed inflected form to determine how to group words, it became necessary to determine where homographs occurred in the corpus, and remove those homographs that seemed the least likely to be employed (see Appendix A for a full listing of items removed). To do this, all the text files were run through AntConc to generate a word frequency list using the lemma list for grouping. All words from the corpus, grouped by inflected forms and not lemmas, were analyzed. Any forms that occurred more than 15 times in the entire corpus of study were verified manually in the lemma list, and where necessary, inflected forms were added or removed from the lemma list to better reflect accurate usage of the language. This helped eliminate infrequently used terms such as the plural noun *sommes* in favor of the far more commonly used second person plural present conjugation of the verb *être*.

The lemma list was additionally trimmed by using the following regular expression to identify hyphenated terms in the lemma list:



```
^.{1,2000}\-[^\>].{1,2000}$
```

These entries were all removed.

2.3.2. CasualTreeTagger

TreeTagger, a free program developed by Helmut Schmid, is a tool for analyzing text files to identify the part of speech (and lemma) of each word in a number of languages, including French. *CasualTreeTagger*, a free Mac OS X front-end for *TreeTagger* written by professor Yasu Imao of Osaka University, allows for multiple files to be analyzed with the same settings in a single batch.

Each individual text was run through *CasualTreeTagger*. This generated a tab-delimited file that indicated the form of each word that occurred, its part of speech, and the lemmatized form. These files were then uploaded to a new database table entitled `tags`, with every entry being tagged for which text it came from and, in the case of non-punctuation items, length in characters (see Fig. 5). 1% of the entries were then manually verified to ascertain how accurate the part-of-speech tagging provided by *CasualTreeTagger* was. This process and the conclusions drawn from it will be discussed in a later section.

```
une         DET:ART   un
petite      ADJ       petit
région      NOM       région
entourée    VER:pper  entourer
de          PRP       de
camps       NOM       camp
retranchés  VER:pper  retrancher
romains     ADJ       romain
...         PUN       ...
tous        PRO:IND   tout
les         DET:ART   le
efforts     NOM       effort
pour        PRP       pour
vaincre     VER:infi  vaincre
ces         PRO:DEM   ce
fiers       ADJ       fier
gaulois     NOM       gaulois
ont         VER:pres  avoir
été         VER:pper  être
inutiles    ADJ       inutile
et          KON       et
césar       NOM       césar
s'          PRO:PER   se
interroge   VER:pres  interroger
...         PUN       ...
quid        PRO       quid
?           SENT      ?
```

Figure 5: Sample sentence from *Astérix le Gaulois* as tagged by *CasualTreeTagger*

2.3.3. Word Count and Average Word/Sentence Length per Album

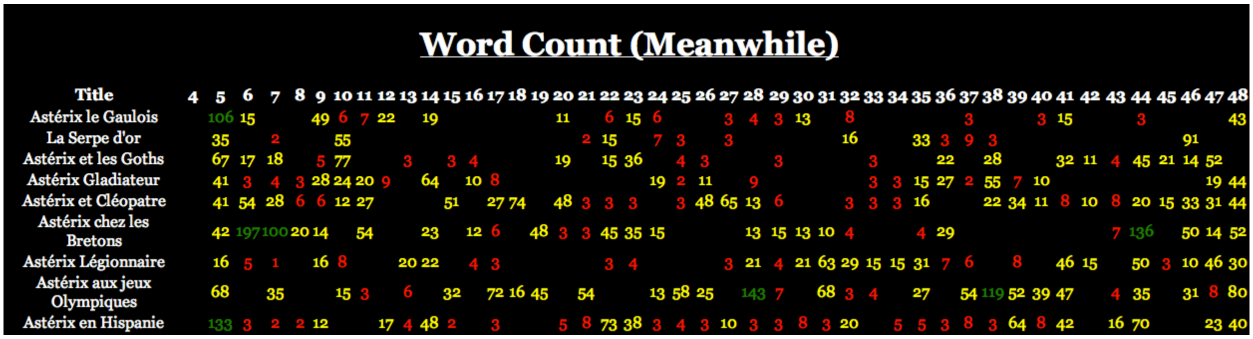
Each album was analyzed for total word count by counting the number of non-punctuation entries in the `tags` table. Average word length for each album was determined by finding the length in characters of each non-punctuation entry, totaling all entries for each album together, and then dividing them by the word count. Average sentence length for each album was determined by taking the word count generated for each album in the `tags` table and dividing it by any entries marked by *CasualTreeTagger* as “SENT”, which are any sentence-ending punctuation marks. Some preparation was required, as the sentence ending punctuation marks (periods, exclamation points, and question marks) often occurred back to back in the *BD* sources (though never in the traditional prose sources). All ‘SENT’ entries in the database were marked as ‘SENT:2’ if the preceding entry was also a sentence-ending punctuation mark. This made the statistics much more accurate for sentence length in the *BD* section of the corpus, as Goscinny and more particularly Uderzo are given to employing multiple such punctuation marks for emphasis. These processes combined into a nested MySQL Query (Fig. 6), the full results of which can be viewed in Appendix B.

```
SELECT a.album, b.words AS words,
        (b.words / a.sentences) AS average_sentence_length,
        c.length AS average_word_length
FROM (SELECT album, count(`id`) AS sentences FROM `tags` WHERE `POS` IN ('SENT')
        GROUP BY `album`) AS a
JOIN (SELECT album, count(`id`) AS words FROM `tags` WHERE `POS` NOT IN ('PUN',
        'PUN:cit', 'SENT', 'SENT:2') GROUP BY `album`) AS b ON a.album = b.album
JOIN (SELECT album, count(form), avg(length) AS length FROM tags GROUP BY album) as c
        ON a.album = c.album;
```

Figure 6: MySQL query used to concatenate word and sentence length results

2.3.4. Word Count per Page (BD Only)

The entry of each *BD* album manually means that every entry was tagged for the page it came from. A simple PHP script totaled the word count of entries occurring on each page of each album, organized by annotations (see Fig. 7). This provides a very quick insight into the structure of the narrative, as pages with more narration boxes and more words (and thus, we assume, plot exposition) should be immediately visible.



Word Count (Meanwhile)

Title	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Astérix le Gaulois		106	15			49	6	7	22		19						11		6	15	6			3	4	3	13		8					3			3	15			3				43
La Serpe d'or		35		2			55											2	15		7	3		3					16			33	3	9	3								91		
Astérix et les Goths		67	17	18		5	77			3		3	4				19		15	36		4	3			3				3			22		28			32	11	4	45	21	14	52	
Astérix Gladiateur		41	3	4	3	28	24	20	9		64		10	8							19	2	11		9					3	3	15	27	2	55	7	10							19	44
Astérix et Cléopatre		41	54	28	6	6	12	27				51		27	74		48	3	3	3		3	48	65	13	6			3	3	3	16			22	34	11	8	10	8	20	15	33	31	44
Astérix chez les Bretons		42	197	100	20	14		54			23		12	6		48	3	3	45	35	15				13	15	13	10	4			4	29							7	136		50	14	52
Astérix Légionnaire		16	5	1		16	8			20	22		4	3					3	4				3	21	4	21	63	29	15	15	31	7	6		8		46	15		50	3	10	46	30
Astérix aux jeux Olympiques		68		35			15	3		6		32		72	16	45		54			13	58	25		143	7		68	3	4		27		54	119	52	39	47		4	35		31	8	80
Astérix en Hispanie		133	3	2	2	12			17	4	48	2		3			5	8	73	38	3	4	3	10	3	3	8	3	20		5	5	3	8	3	64	8	42		16	70			23	40

Figure 7: PHP tool to calculate page composition statistics by category

2.3.5. Word Lists, Keyword Lists, and N-grams

AntConc provided an effective solution for generating word lists, generating keyword lists to identify those terms that were more prevalent in one section of the corpus than they are in another section of the corpus, and in identifying N-grams. Each text was analyzed individually to determine how many different words were used in it (using the UIMA lemma list), and then analyzed again for how many unique forms were used (without the UIMA lemma list). Then, each collection of texts was analyzed as a whole for the most frequently used words, the most frequently used N-grams (for values of N between 3 and 8), and the comparative keywords in each collection as compared to each other collection.

2.3.6. Part of Speech Analysis

CasualTreeTagger was employed for a batch analysis of all the text, separated into individual files. Of the 411,720 entries generated (including punctuation marks), approximately 14,000 entries weren't recognized. Of these unrecognized entries, over 3,500 of these entries were proper nouns, and many were forms that weren't recognized by *TreeTagger* due to characters not matching the expected case: proper nouns in lower case, demonstrative adjectives that were capitalized, etc. Most notably, these occurred in *Astérix*, where all text is in a single case.

After manually fixing any entries with more than 25 occurrences, only 1.7% of the entries remained marked as “<unknown>”. It would also appear that TreeTagger has a tendency to mark any unrecognized words as a noun, as over 78 % of “<unknown>” forms were tagged as either nouns or proper nouns. Though many of the entries were indeed proper nouns, it seldom chose correctly outside of the names of major characters in the series such as Astérix, Obélix, and Panoramix. In those cases, it only benefitted by the fact that those particular names are used often, as they occurred more frequently than other entries that were mislabeled.

2.4. Data Correction

With all of the part of speech tags in place (barring the remaining “<unknown>” entries), specific manual effort was undertaken to verify how accurately *CasualTreeTagger* had identified each word. 4,056 non-punctuation entries were verified from the corpus at random (approximately 1% of the corpus), with the following results:

Table 3: Data correction for automated part-of-speech tagging

	Entries Checked	Incorrect	Percentage Incorrect
Astérix by Goscinny	1050	59	5.62%
Astérix by Uderzo	650	41	6.31%
Le petit Nicolas	1185	15	1.27%
Tintin	80	7	8.75%
Yoko Tsuno	90	10	11.11%
Le petit prince	184	3	1.63%
Le grand meaulnes	817	20	2.45%
TOTAL	**4056**	**155**	**3.82%**

The distribution of entries checked is representative of the number of words in each section of the corpus. Looking at the percentages, it is immediately evident that the TreeTagger engine far more accurately identifies traditional prose than it does *BD*.

The additional difficulties posed by the *bandes dessinées* can mostly be grouped into a few cases. Firstly, a good number of interjections and onomatopoeias were incorrectly identified as nouns or proper nouns. Secondly, in *Astérix*, the fact that the text is all in capitals means that when the text was analyzed, it was converted to lower case. It had difficulty discerning between nouns and proper nouns, likely because the proper nouns weren't capitalized as would be expected by convention. Thirdly, the *bandes dessinées* had a number of verbs conjugated in the imperative, however TreeTagger identified them as the present tense. By far, however, the most difficulty was posed by adjectives. Any adjectives that are identical to the past participle of a verb were identified as a past participle. At a grammatical and etymological level, however, these two are related, as one derives from the other.

In the second and third cases, these groupings are obviously close to the correct groupings, slipping from a proper noun to a common noun and from the imperative to the present tense. In all but the first case, if the confused categories are either grouped together or ignored entirely, it should minimize any widespread misidentification.

In more isolated cases, words that were in non-standard forms tripped up *TreeTagger*, for example those words that started with an asterisk symbol. One area of potential concern, depending on the level of analysis desired, is the fact that *TreeTagger* tends to group many related parts of speech into a single category. *De* is always tagged as a preposition, though it can at times act as a partitive or an indefinite article. All demonstratives were grouped together, though it doesn't differentiate between demonstrative adjectives and demonstrative pronouns. It also had difficulty with distinguishing the pronoun *en* from the preposition *en*.

On the other hand, however, the automatic part of speech tagging succeeded in many areas that could have proved potentially difficult. It did an excellent job with homographs; for example, *livres* was correctly identified as a present-tense conjugation of the verb *livrer* as opposed to the plural inflection of the noun *livre*. It was able to determine when the word *si* acted as an adverb and when it acted as a conjunction, and did a good, though not flawless, job of determining instances of *la* where it functioned as a personal pronoun as opposed to a definite article, such as in "*sans même la regarder partir*."

In the end, if we take the most common errors into account, the text seems to be reasonably well tagged. However, specific precautions must be taken to minimize the impact of words that are more likely to be misidentified for any of the above-mentioned reasons.

2.5. Resulting Data

The following sections will be dedicated to the analysis of the data generated, though only selections of data will be presented for discussion. For reference, full results are presented in the Appendices. The data can also be accessed in an electronic format. Please see Appendix F for further details and limitations.

3. COMPARISONS

3.1. The Comparative Complexity of Language Used

Operating under the assumption that the presence of longer words and longer sentences corresponds to more complicated language use, averages for both of these values were compared. In Table 4, we have a summary of the results of this information. For the full dataset, see Appendix B.

Table 4: Summary of Word and Sentence count analysis.

Work	Total Words	Average Sentence Length in Words	Average Word Length in Letters	Relative Pronouns	Sentences per Relative Pronoun
Le petit Nicolas (average)	1440	17.76	3.85	31	2.83
Le petit Prince	15456	11.56	4.072	257	5.20
Le grand Meaulnes	67922	20.61	4.41	1459	2.26
Astérix (Goscinny, average)	6321	6.94	4.30	103	9.07
Astérix (Uderzo, average)	6487	6.95	4.33	121	7.75
Tintin au Congo	6740	6.71	4.20	97	10.35
Trio de l'étrange	7589	8.75	4.54	88	9.85

We can quickly see that the traditional prose works have longer sentences and employ more relative pronouns than their *BD* counterparts. It is evident that these traditional prose works lend themselves to very different sentence structures than *BD*. The 69 stories from *Le petit*

Nicolas have an average of 17.76 words per sentence. By comparison, Goscinny's *Astérix* only employs 6.94 words per sentence, and Uderzo's *Astérix* is very similar at 6.95 words per sentence. The reference works bear this difference out, as *Le petit prince* has a words-per-sentence value of 11.56, *Le grand Meaulnes* a value of 20.61, *Tintin* a value of 6.71 and *Yoko Tsuno* a value of 8.75.

To verify whether the structure of the sentences themselves was more or less complex, a count of relative pronouns was undertaken. The count of relative pronouns was then used to calculate a value indicative of how many sentences would occur between the use of a relative pronoun. This was achieved by dividing the number of sentences by the number of relative pronouns employed. On average, *Le petit Nicolas* uses a relative pronoun every 2.83 sentences, *Astérix* by Goscinny every 9.07 sentences, and *Astérix* by Uderzo every 7.75 sentences. So, the critique of less elegant wordplay on the part of Uderzo wouldn't appear to have a direct relationship to the complexity of sentences as measured in relative pronouns.

Also, while the original calculations for sentence length were perfectly appropriate for the prose, the *BD* gave uncharacteristically low words per sentence lengths, far beyond what could have been anticipated. It became necessary to create a new Part of Speech tag in the database for subsequent sentence-ending punctuation. Sentences ending in three exclamation points, for example, were being counted as three sentence ends (see section 2.3.3 above for further details). Correcting this problem revealed two contrasts between the works. First, that the prose didn't employ any multiple punctuation at all, where as it was very frequent in every *BD* album. Second, Uderzo uses far more multiple punctuation than the other *BD* authors. Where Goscinny used an average of 103 superfluous punctuation marks in each album, Uderzo employed an average of 186.5, almost twice as many. By way of comparison, Hergé only used 52 superfluous punctuation marks, and Leloup used 111.

As a beginning to our analysis, the numbers above reveal some very distinct differences between prose and *BD*, however, Goscinny's *Astérix* and Uderzo's *Astérix* look very similar when merely looking at averages. The difference in average number of words employed accounts for the equivalent of about 1 extra page of text on Uderzo's part, or an extra 2.2% worth of text per page. The only discernable differences visible from this point are that Uderzo uses more relative pronouns and more punctuation marks than Goscinny.

3.2. The Comparative Lexical Richness of Titles

Each work had a word list generated with *AntConc*, and using a lemma list, we are able to accurately see how many different words (lemmas) are used, not just how many different *forms* are used. An interesting point of comparison immediately arises when looking at a summary of this data (see Table 5), in that each individual work in the *Le petit Nicolas* series contains a higher percentage of unique lemmas than either of the authors of *Astérix*, however, the total number of unique lemmas in each *series* paints a different picture.

Upon closer inspection, we can easily explain this disparity. Though *Le petit Nicolas* is more lexically diverse within the context of a single story, the entire body of Uderzo's *Astérix* has more unique lemmas, despite being only half its size in number of total words. The elevated use of unique lemmas in any given work of *Le petit Nicolas* can be partially attributed to the fact that each story is quite short, at only about a quarter of the length in words as the average album of *Astérix*. Logic dictates that as an author continues to write, he will have access to fewer and fewer words and expressions that he hasn't already used; the longer one writes the more one is forced to reuse idiomatic expressions and grammatical structures.

Table 5: Summary of Unique Lemma Analysis.

Title / Series	Unique Lemmas Used	Total Words	Percent Unique
Le petit Nicolas (average per work)	316	1458	21.7 %
Le petit Prince	1737	16409	10.6 %
Le grand Meaulnes	4695	72286	6.5 %
Astérix (Goscinny) (average per work)	1137	6262	18.2 %
Astérix (Uderzo) (average per work)	1245	6426	19.4 %
Tintin au Congo	1254	6877	18.2 %
Trio de l'étrange	1743	8099	21.5 %
Le petit Nicolas (whole series)	3109	99168	3.1 %
Astérix (Goscinny) (whole series)	5718	87673	6.5 %
Astérix (Uderzo) (whole series)	4555	51404	8.7 %

What can be demonstrated here is that two albums of *Astérix* are less lexically similar than two stories of *Le petit Nicolas*. In this sense, word-per-word, Goscinny's writing is lexically richer in the pages of *Astérix*.

Similarly, we can see that Uderzo's writing is lexically richer than Goscinny's writing. There are a few possible explanations for this. Firstly, it is conceivable that Uderzo wanted to send the characters into new territories, both literally and literarily, as approximately two out of every three of Goscinny's stories had been set in and around the village where Astérix lives. Indeed, Uderzo sees Astérix and company visiting India, going on an Odyssey, touring the world, and visiting with aliens from Space. Secondly, perhaps Uderzo employs more invented proper nouns than Goscinny did. Thirdly, it is also possible that Uderzo uses fewer repeated expressions.

As we look at keywords and N-grams in the next two sections, we will be able to better address these possibilities.

3.3. Keywords

Comparisons of the word lists were made using AntConc. Of particular interest are the authors' uses of proper names, as Goscinny has made a reputation for *Astérix* with the inclusion of many creative and humorous names. Looking at the top ten ranked entries (see Table 6), we see that in Astérix, the keyword *être* stands out, while in *Le petit Nicolas* the verb *avoir* is more prevalent. Both authors' *Astérix* have very similar keywords, while comparing *Nicolas* to *Astérix* elicits the prominence of Nicolas' parents (*papa*, *maman*) and the setting and characters of ancient Gaul (*romain*, *gaulois*, and *Obélix*).

Expanding our scale to the top 250 entries (see Appendices), the comparison between *Le petit Nicolas* and *Astérix* continues to reveal a large number of the main characters' names prominently in the keyword lists. Of more interest, however, is the prevalence of *avoir* in *Le petit Nicolas*. This is probably due to grammatical structures, as *Le petit Nicolas* employs the past tense more frequently than the *BD*, as it is a part of the supposed narrator's style in retelling his journeys and exploits.

If we make a similar comparison of Goscinny and Uderzo, we notice a difference in the usage of characters' names. Goscinny's *Astérix* employs invented proper names less frequently than Uderzo's *Astérix*. In looking at the 250 most prominent entries, Uderzo features 20 invented proper names, the topmost entries of which come in at #41 and 42. Only 15 invented proper names occur in Goscinny's 250 most prominent entries, the topmost are ranked #53 and 111. This difference in usage might begin to explain a small part of the difference in the number of lemmas occurring in each author's works.

Table 6: Top Ten Keywords

Goscinny to Uderzo					**Uderzo to Goscinny**			
Rank	Freq	Keyness	Keyword		Rank	Freq	Keyness	Keyword
1	3235	2433.398	être		1	1862	2741.317	être
2	1956	1334.666	avoir		2	1100	1688.439	avoir
3	1391	1282.873	les		3	637	1268.637	les
4	3859	737.152	le		4	2314	814.621	le
5	1029	725.639	aller		5	1394	719.777	ce
6	2219	545.02	ce		6	440	506.413	aller
7	1395	278.724	du		7	245	387.075	pouvoir
8	236	217.655	falloir		8	127	242.157	vouloir
9	329	197.702	pouvoir		9	118	216.507	devoir
10	255	184.594	vouloir		10	687	208.601	du
Nicolas to Astérix					**Astérix to Nicolas**			
Rank	Freq	Keyness	Keyword		Rank	Freq	Keyness	Keyword
1	6038	6938.327	avoir		1	3235	3955.585	être
2	3967	3923.108	être		2	1956	2437.706	avoir
3	5442	1814.661	le		3	1391	2103.318	les
4	1665	1363.64	dit		4	2219	1424.042	ce
5	1038	1316.127	les		5	1029	975.579	aller
6	810	934.144	papa		6	1395	792.384	du
7	3541	797.108	et		7	457	691.025	romain
8	668	725.447	pouvoir		8	3859	647.253	le
9	575	705.441	maman		9	340	501.718	gaulois
10	793	594.907	aller		10	308	465.724	obélix

Specifically, Table 6 shows the most commonly used words in each section of the corpus, giving additional emphasis to those words that aren't used in a reference corpus (in this case, calculated by means of a log-likelihood ratio test). Comparing Goscinny to Uderzo, for example, shows that the most frequently used words in both authors' versions of Astérix are *être*, *avoir*, *les*, and *le*. However, comparing Goscinny's *Nicolas* to his *Astérix* shows that though *avoir* and *être* are still the two most commonly used words, words such as *papa* and *maman* are disproportionately more common.

3.4. N-grams

The N-gram results for *Le petit Nicolas* expose to what extent the narrative style of the work differs from *Astérix*. N-grams were counted for values of N between 3 and 8, where a phrase occurred at least three times. Of the top 20 entries for *Nicolas*, only 2 don't feature the *passé composé*, while all of the top 20 entries for *Astérix* that feature verbs are conjugated in the present tense. This supports the reasoning deduced above of the importance of *avoir* in *Le petit Nicolas*.

There are many similarities between Goscinny and Uderzo's prose in *Astérix*. For example, both authors feature expressions like *de la potion magique* and *de potion magique* in the top 20 N-gram entries. Similarly, the famous expression *ils sont fous* comes in at #43 for Uderzo and #41 for Goscinny. However, it seems as though derivations of *ils sont fous ces romains* is about the extent of the fixed expressions that Uderzo was able to borrow from his former collaborator.

When N is set to 8, Uderzo has only one entry, the repeated sound effect *clap*, however, Goscinny's *Astérix* has 9 entries, including formulaic expressions such as *le petit village gaulois que nous connaissons bien* and *la 1ère legion 3ème cohorte 2ème manipule 1ère centurie*. Such expressions seem reminiscent of expressions from *Le petit Nicolas*, for example the epithet applied to Agnan, that he was *le premier de la classe et le chouchou de la maîtresse*. An additional difference is in the fact that though Uderzo is likely to repeat a particular word or phrase again and again within the confines of a single album (*de l'huile de roche*, kiçàh, and many invented names, for example), Goscinny continued to use things in later albums and stories again and again. Agnan's epithet, for example, is used in 20 different stories in the corpus. The phrase *le petit village gaulois que nous connaissons bien* appears in four separate volumes of

Astérix (as penned by Goscinny) in the corpus. Even *la 1ère legion 3ème cohorte 2ème manipule 1ère centurie* appears in two separate albums in the corpus (*Astérix légionnaire* and *Astérix chez les belges*), published 12 years apart. It is evident from these lists that Goscinny favored certain expressions, which he enjoyed bringing back.

In fact, after having come to this conclusion, a cursory re-reading of the Goscinny biography by Guillaume and Bocquet reveals that the authors identify heavily their positive associations with these epithets and commonly used phrases. Indeed, their first sentence is "Tout le monde connaît Astérix, Obélix, et la fameuse "1re légion, 3e cohort, 2e manipule, 1re centurie" de vaillants Romains…" (Guillaume and Bocquet 13) In the chapter where *Nicolas* is covered, they go on to introduce each character by those things that the characters are known for, such as Alceste being a "grand amateur de sandwiches" and Agnan being "le premier de la classe." (Guillaume and Bocquet 97)

This repetition of phrases may also partially account for the reduced number of lemmas used in each series (compared to Uderzo's work, that is). This could have influenced Norot's comments that after the death of Goscinny, the "language" and "jokes" were gone. It seems reasonable to assume that a seasoned reader of *Astérix* would be expecting to see verbal references back to the previous works, and instead found only occasional footnotes specifically identifying the albums that are being referenced as, for example, Astérix and his compatriots are flying over foreign lands previously visited on the way to India in *Astérix chez Rahàzade.*

One other difference that can be noted from the comparison of N-grams is that the structure and usage of narration boxes in the two authors' *Astérix* may be different. The expression *pendant ce temps* figures as #14 on the list of N-grams for Goscinny (4.1 times per volumes), but only #30 for Uderzo (2.5 times per volume). Though no N-grams were analyzed for values of N as 2, searching the `corpus` table of the database manually for expressions contained in narration boxes reveals that Goscinny also favored *peu après* and *plus tard* in narration boxes. *Peu après* is used an average of 7 times per volume by Goscinny, but only an average of 4.25 times a volume by Uderzo. On the other hand, *plus tard* is used 5 times a volume by Uderzo, where Goscinny had only used it about 2.6 times a volume.

The usage of these and other phrases becomes important in regarding the structural composition of the two authors' works. The structure of albums as viewed through the distribution and length of narration boxes will be developed in section 3.6.2.

3.5. The Comparative Distribution of Parts of Speech

Following the data correction, a full count of each works' parts of speech was made using a PHP script, the full results of which are viewable in Appendix D. Bearing in mind the difficulties encountered by *CasualTreeTagger* in distinguishing between certain parts of speech, and for the purposes of not over complicating things, all nouns are grouped together (both common and proper) and certain parts of speech are not being considered as individual entities.

A Chi-square test was performed on the data for each individual work. One limitation of the Chi-square test is that it assumes that data points aren't influenced by other data points (Larson-Hall 241). Because we're looking at only a *count* of words categorized by parts of speech, we're going to look at the data under the assumption that though there are rules that force words into a specific order, there is nothing that forces an author to use any given word and nothing that influences their distribution. Authors are free to choose the words and terminology that they like, and that an author has some leeway in choosing even those words that are included based upon the use of another word, such as demonstratives and articles (*le porc*, *mon porc*, *ce porc*, etc). The Chi-square test is advantageous in these comparisons because the samples don't need to be the same size. It can compare works of different length, which allows for comparison between the shorter *Nicolas* stories, the medium-sized *BD* and the longer reference novels. Unlike parametric analyses, there is no necessity to assure that the data is distributed normally. However, as we're looking at patterns in writing, it is important to note that the distribution of parts of speech followed relatively normal curves, though those curves had varying degrees of skewness and kurtosis. In *Le petit Nicolas*, for example, the presence of verbs conjugated in the present tense had a skewness value of -.495 and a kurtosis value of -.104. This relatively normal (if skewed) distribution is important because it indicates that there are indeed patterns in the authors' writings, and that one *Nicolas* story has a similar composition to another *Nicolas* story.

In Figure 8, we get a quick idea of the distribution of present tense verbs. No parametric tests on the data would be possible, however, as the numbers for each group are skewed in opposite directions. Unfortunately, this causes difficulties in generating confidence intervals, as the skewness and kurtosis values are not uniform across the part of speech categories. Because confidence intervals require a normal distribution, they were not generated for this data. As an example, for *Nicolas*, the present tense has a skewness value of -.495, the conditional has a value of 1.099, the future has a value of .669, the imperfect has a value of .251, and the ensemble of the remaining verbs has a value -.097. As such, no uniform set of transformations would be able to normalize the data.

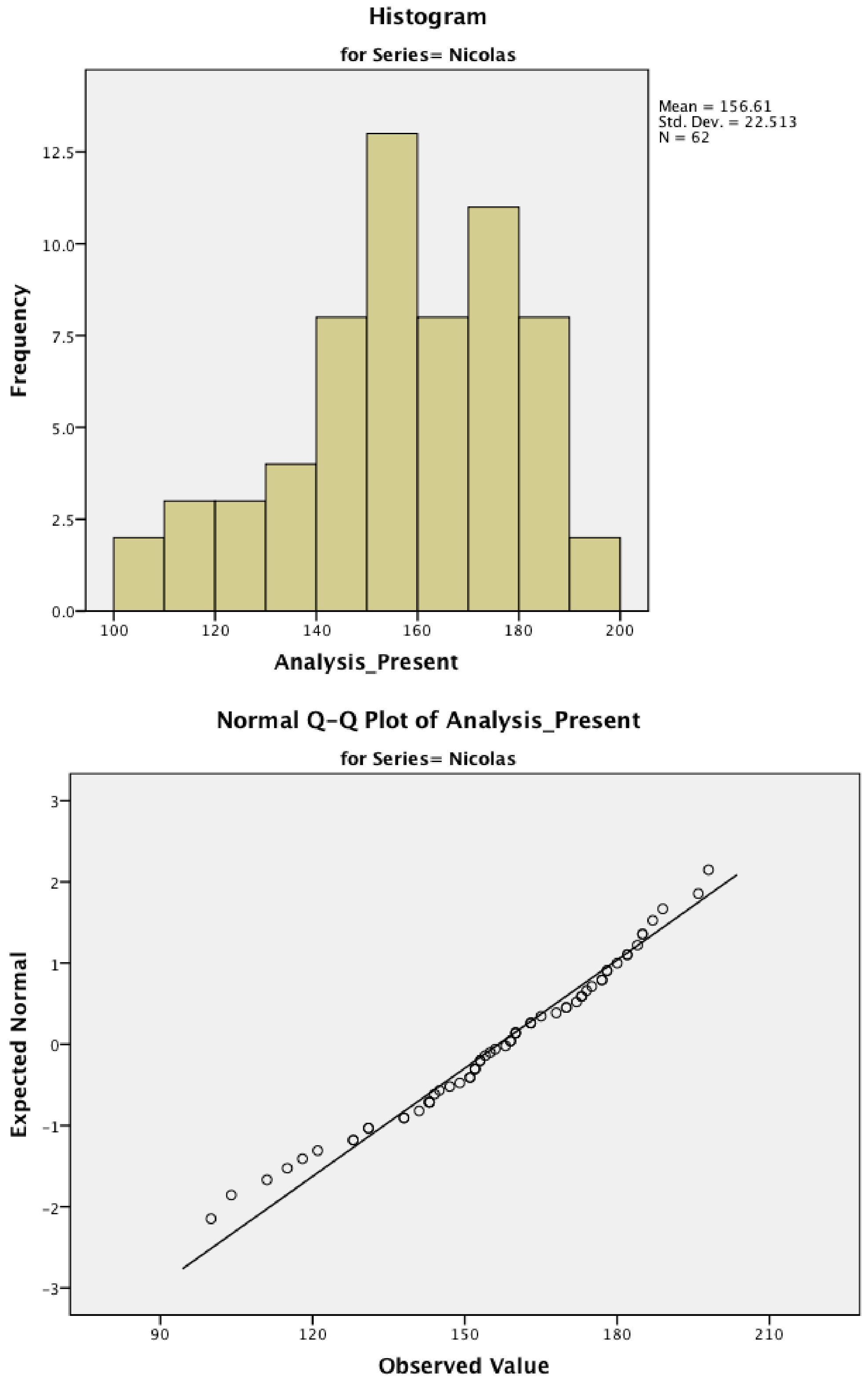


Figure 8: Graphical representation of the distribution of present tense verbs in *Le petit Nicolas,* where the x-axis represent number of occurrences

Running a power analysis in *R*, a comparison of two data points on the part of speech data for each individual work gives us a power of 0.9965717 to discover a large effect (if one exists), a power of 0.7990557 to discover a medium effect, and a power of 0.09759641 to discover a small effect.[2] As a result, any effects seen are very likely to really exist at a medium to large level, as opposed to merely being a statistical aberration. Because additional degrees of freedom increase the chances of getting a falsely significant comparison, all of the comparisons undertaken using a Chi-square test are limited to one degree of freedom (i.e. only two figures are being compared between the two works at any given time). Though Larson-Hall believes that a significance level of .1 is probably sufficient, all numbers were calculated for a more traditional significance level of .05. Subdividing each volume of Nicolas into its individual stories, we have 87 texts that can be tested. As a result, for each test, the data was run through a series of 2,964 pairwise Chi-square tests, and then averages were taken comparing each of the 87 texts to a body on the whole to help see where patterns occur.

To aide in the identification of where patterns occurred, an automated spreadsheet was created which would compare the Chi-square values for each of the 2,964 comparisons in a large grid. Through the use of conditional formatting, all values that showed a result of less than .05 were marked in maroon, while all values of .05 and above were marked by gradations of orange (for low values) to green (for high values). The resulting "heat map" could then be shrunk to show patterns among all of the works at a single glance. Figure 9 is an example of the comparison of the count of verbs conjugated in the present and imperative with a count of all the remaining verbs in a text (the data of which is analyzed in section 3.5.3.).

[2] Three tests were run using the pwr.chisq.test function with values of N (number of observations) = 87, DF (degrees of freedom) = 1, sig.level (significance) = .05, and w (effect size) = .5, .3 and .1 respectively.

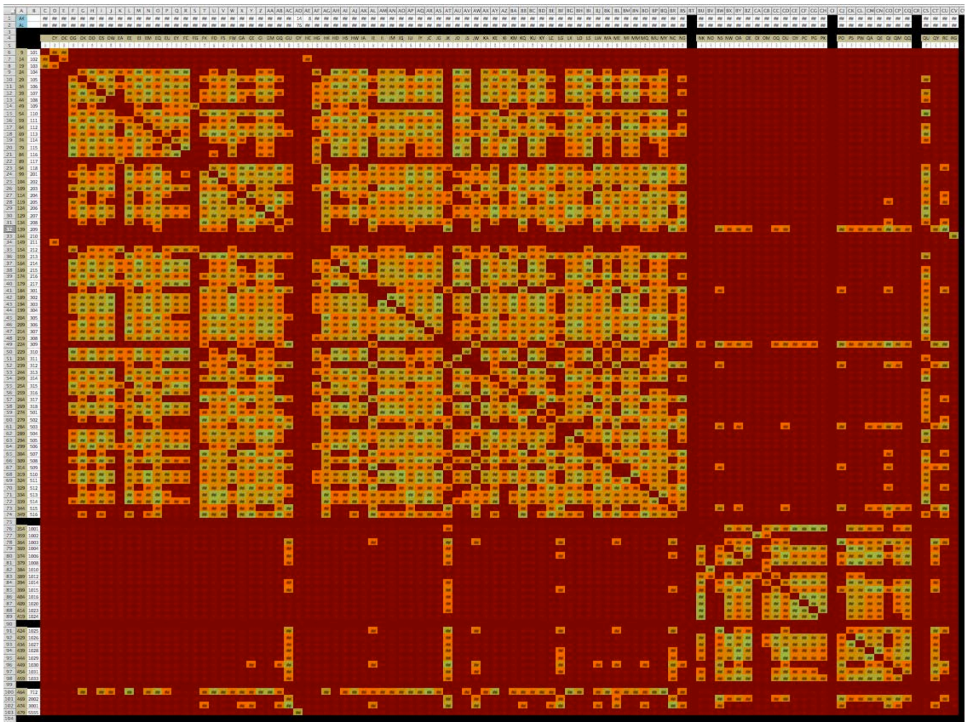


Figure 9: A heat map where cells have been color-coded to show Chi-square results; this particular heat map compares occurrences of the present and imperative to other verbs

In Figure 9, the first two thirds of the columns and rows correspond to the *Nicolas* stories, while the second grouping corresponds to Goscinny's *Astérix*, the third grouping to Uderzo's *Astérix*, and the last grouping to the reference works. The black notches across the sides help delineate the groups. The data is presented twice in this chart so that averages can be taken across rows and columns. The diagonal line from the top left corner to the bottom right corner are blank cells where a text would have otherwise been compared to itself. As such, the results are reflected across this diagonal.

The grouping of orange and green cells for the top left two thirds of Figure 9 shows that the usage of the present and imperative in most of *Le petit Nicolas* is very similar to other stories within that same body. The large maroon sections beneath it and to its right indicate that the usage of these tenses is below our .05 threshhold for similarity. The grouping toward the bottom right corner shows that the various volumes of *Astérix* by both authors. The first entry in the last section (the reference texts), are mostly maroon, though in a standout exception, the results for *Le petit prince* show a number of similarities with the text of *Le petit Nicolas*; this is represented by the very visible (if patchy) stripe that is 4 entries from the bottom and the reflected stripe 4 entries from the right side of the figure. These sorts of visual patterns are exactly the advantage of such heat maps, as it is often easier for the human eye to spot visual patterns than it is to identify numerical trends.

3.5.1. The Comparative Use of Nouns, Pronouns, and Demonstratives

Due to the graphical nature of the *BD*, many questions arise about the economy of words that can be introduced due to the visual demonstration of information. Specifically, it seems reasonable to assume that proper nouns will need to be used less frequently in *BD*, as readers will see the individuals being referred to. As a result, it would also stand to reason that demonstratives and pronouns might be used in a higher proportion than in traditional texts. Because of the above-mentioned difficulties in *TreeTagger* correctly distinguishing between common and proper nouns, the comparison was performed twice, once looking at only proper nouns, and once looking at all nouns.

In comparing only the proper nouns (see Fig. 10), it seems that the numbers can't justify any broad claims. The average correspondence between Goscinny and Uderzo is 1.73e-01, though no real tendencies arise, as the numbers don't really vary that much between *Astérix* and *Nicolas*, even when comparing all nouns. The levels of correspondence between *Tintin* and both authors of *Astérix* are 1.03e-05 and 4.80e-09 for Goscinny and Uderzo respectively for proper nouns, and remained below the .05 threshold when looking at all nouns. The level of correspondence between *Tintin* and *Yoko Tsuno* is even smaller, at 3.05e-15, and LeLoup had actually worked for Hergé at the Studios Hergé for almost 17 years. In any case, it seems that even though LeLoup adopted Hergé's *ligne claire* style, it seems his textual composition is his own. In looking at all of the results by numbers, the standout work was *Le petit prince* which is visible as the first entry in the last grouping. The average Chi-square result comparing it to all the other works was 1.20E-07. Having read *Le petit prince*, this makes sense, as there are very few proper nouns. Most of the characters in the work are identified by titles and epithets such as *la rose*, *l'homme d'affaires*, or *le renard*.

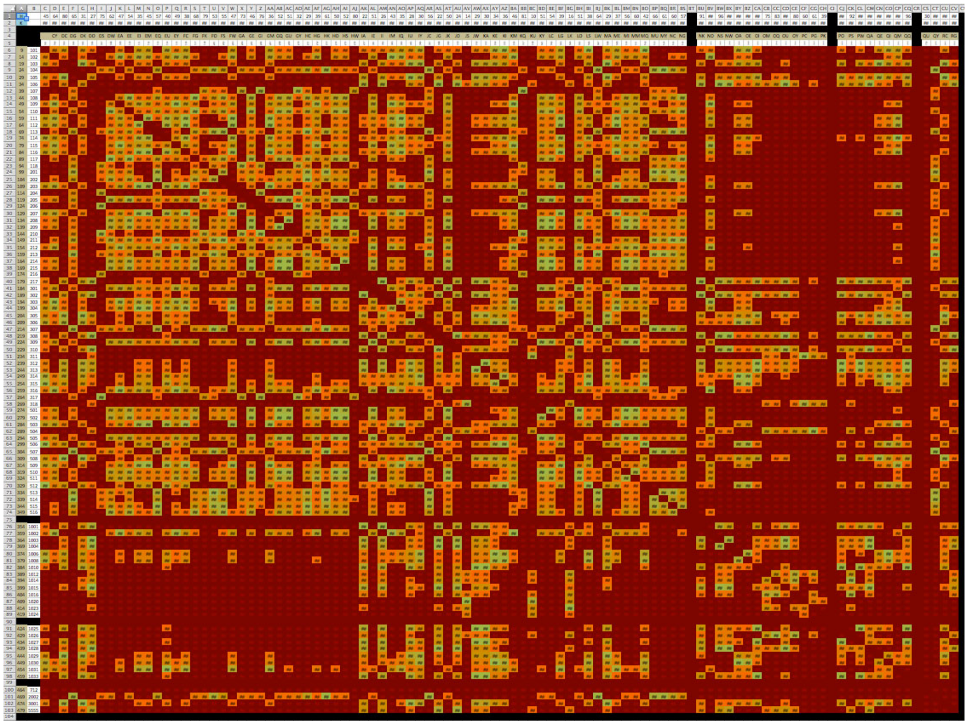


Figure 10: Heat map for proper Nouns

When expanding to look at all nouns (see Fig. 11), however, the result for *Le petit prince* jumps startlingly to an average of 7.70e-01. This could be attributed to *TreeTagger* incorrectly identifying the proper nouns in the smaller sections of the corpus (i.e. the reference works), because the only manually corrected entries were those occurring more than 25 times, however, a glance at the heat map for the comparison shows that almost all of the texts corresponded very well with each other across all nouns. The stories that stand out the most are the first three Nicolas stories (in the top left corner), and a trilogy of stories centering on a soccer match (a visible gap 25% of the way from the top corner). These will be discussed in further detail later.

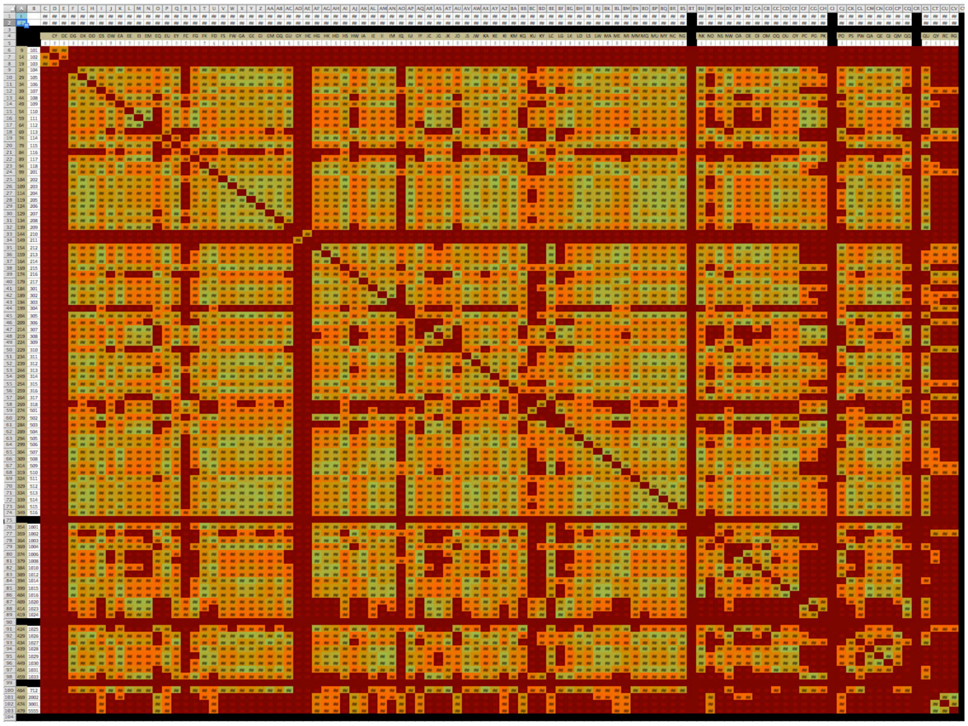


Figure 11: Heat map for noun usage

The expected correlation of increased use of demonstratives and pronouns, however, showed even less promise than the previous assumption (see Fig. 12). Though selected works stood out (such as the first three *Nicolas* stories, and the soccer trilogy, where again the only discernable similarity in results was amongst themselves), no trends emerged, and the numbers comparing traditional texts and *BD* showed no discernable pattern of difference than those comparing *BD* and *BD* or traditional text and traditional text.

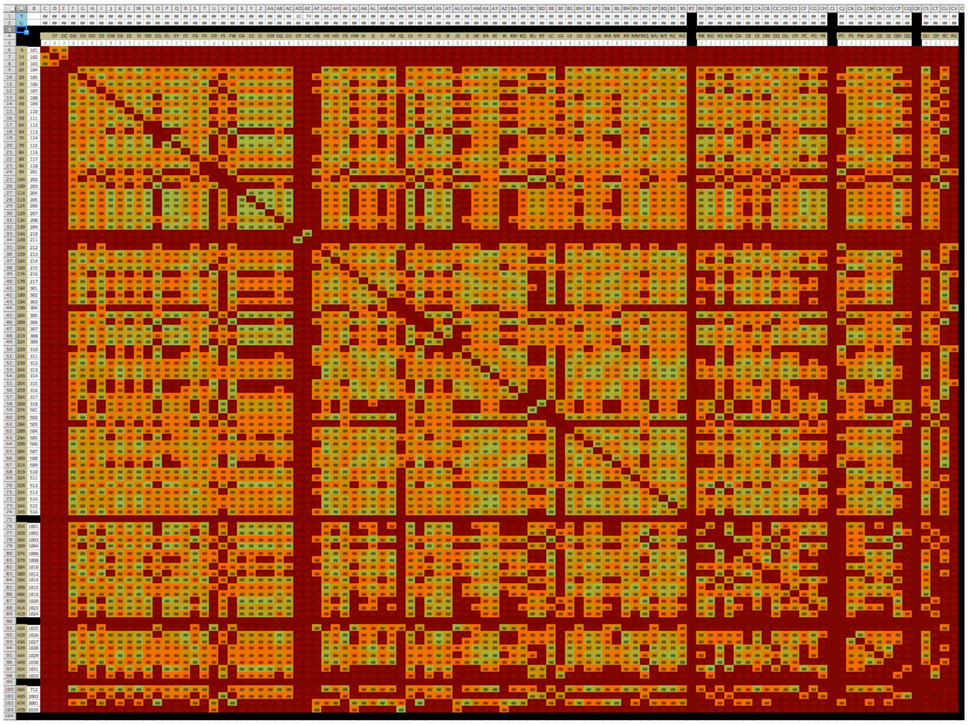


Figure 12: Heat map for pronoun usage

3.5.2. The Comparative Use of Adjectives

Similar to the hypothesis above concerning proper noun and pronoun usage, it also seemed reasonable to assume that fewer descriptive words would be used in the *BD* in general, seeing as the visual elements of each work could provide a better economy of space for portraying such characteristics. The pair-wise comparisons showed a number of interesting points. For example, though I anticipated that the adjective usage in *BD* wouldn't correspond very well with more traditional texts, it seems that the stories from *Le petit Nicolas* are far more different as a body than the other works. Interestingly, though, a handful of stories buck the trend of the other stories in their respective volumes, and correspond more closely to the adjective usage in *BD*. In particular, three stories that appear back-to-back in *Les récrés du petit Nicolas*, specifically *Le football*, *1re mi-temps*, and *2e mi-temps*, return the following chi-square values:

Table 7: Adjective usage in selected stories from *Les récré du petit Nicolas*

	Compared to average *Nicolas*	Average *Astérix* by Goscinny	Average *Astérix* by Uderzo	Reference Literature	Reference *BD*
Le football	0.038	0.570	0.587	0.666	0.807
1re mi-temps	0.006	0.232	0.237	0.179	0.072
2e mi-temps	0.143	0.456	0.406	0.280	0.574

These three stories form a short trilogy, the second and third of which are not written in the style of the other *Nicolas* stories, but instead are narrated in the third person. The first story is also slightly different from other *Nicolas* stories in that, though Nicolas is narrating as normal, it seems less Nicolas-centric in that the focus is on a whole team, and Nicolas proffers fewer commentaries on the events. It's more matter of fact, in a way. These same three stories also had a stark contrast from the other *Nicolas* stories when all nouns and pronouns were compared, though they didn't correspond particularly closely to any of the *BD* works in those instances.

3.5.3. The Comparative Use of Verbs

Beyond the already demonstrated predilection for Goscinny's use of *passé composé* with *Le petit Nicolas* and the present with *Astérix*, the part of speech tagging data was used to determine to what extent the verb tenses used differed between works.

For this particular test, the imperative and the present tenses were combined, and compared to the use of all other verbs. Though the *passé composé* is one of the most frequently used tenses, the past participle wasn't analyzed by itself for two reasons. Firstly, owing to *TreeTagger*'s difficulties in discerning between it and adjectives, and secondly because it always occurs with another verb, but we didn't generate any collocational data which indicates to us whether that past participle was used to form the *passé composé*, the *futur antérieur*, or any other compound tense. The Chi-square results (visualized in Fig. 9 above) again show that the works which most stood out were the first three *Nicolas* stories, and the soccer trilogy. However, perhaps nowhere in the results comparing parts of speech is there a clearer delineation that *Nicolas* corresponds to *Nicolas,* but none of the other works. The *Nicolas* stories almost all showed a result of less than .05 when compared to the *BD*, and most of the Nicolas stories compared favorably to *Le Petit Prince*.

In comparing other verb tenses, limitations arose due to the fact that many of the texts didn't use the *passé simple* or the subjunctive (either present or imperfect). Comparisons were made for the conditional (see Fig. 13), the infinitive (see Fig. 14), and the imperfect (see Fig. 15). The conditional's results showed no distinct patterns either visually or numerically. The use of the infinitive, however, showed the grouping of *Nicolas* and everything else as separate categories, though less distinctly than the present tense. The imperfect, more than any previous comparison, shows exactly *how* the first three *Nicolas* stories differ from the others. Their usage of the imperfect tense much more closely corresponds to most of the *BD* in the corpus, including *Tintin* and *Yoko Tsuno*.

The fact that the difference can be so much more clearly seen in the present, the infinitive, and the imperfect, is perhaps due to the fact that they represent 46%, 15%, and 10% of all verbs in the corpus, respectively. The conditional (and the future, not included for this express reason) only represent 2% each of the verbs in the corpus.

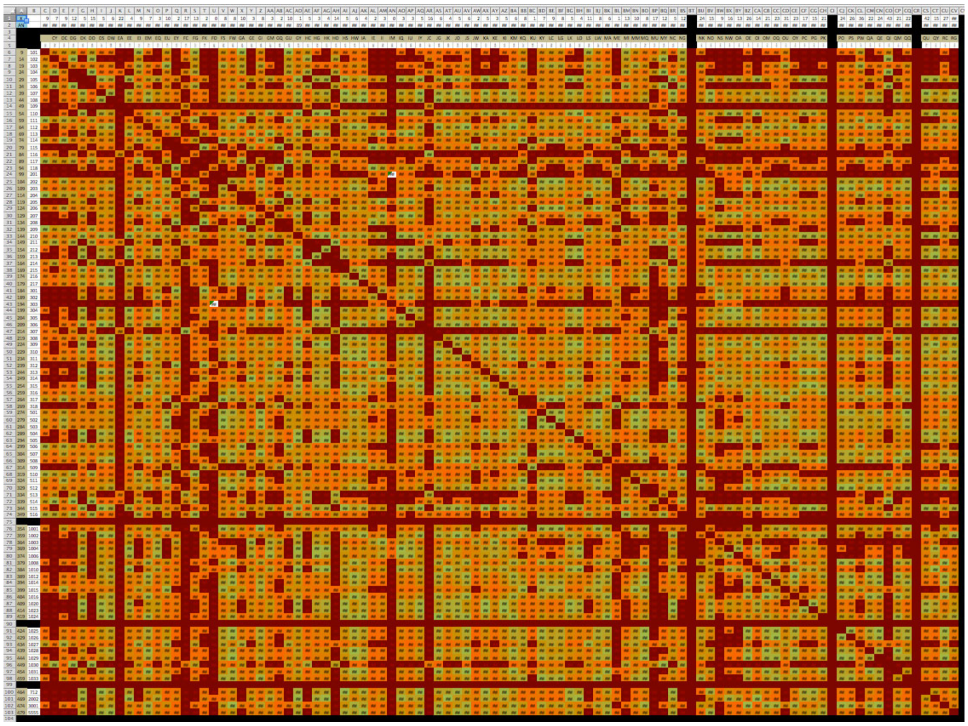


Figure 13: Heat map of conditional verbs

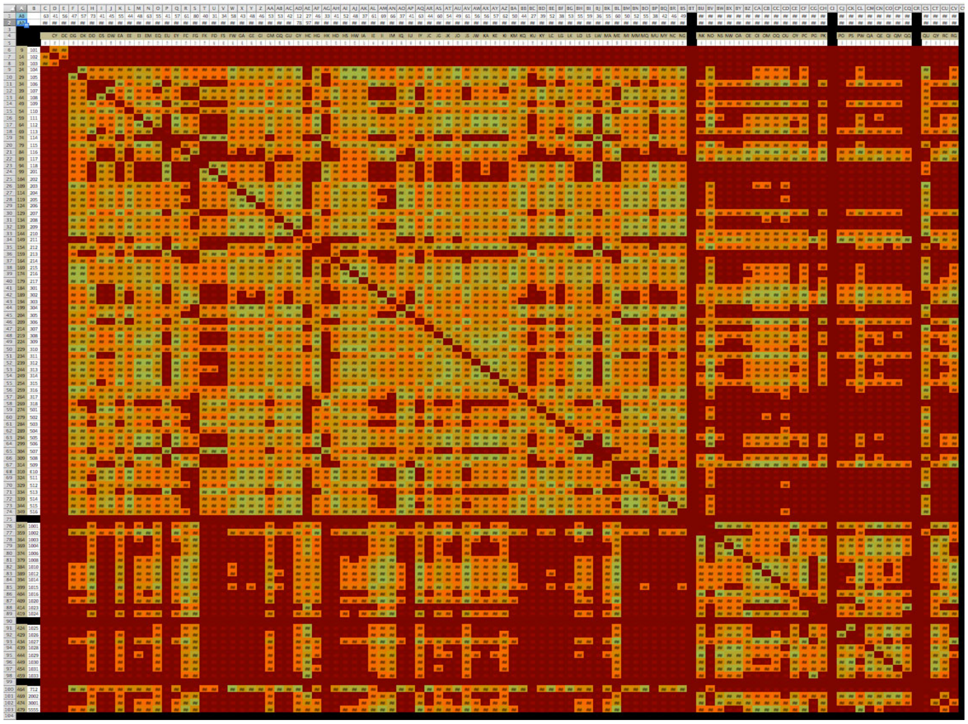


Figure 14: Heat map of infinitive Verbs

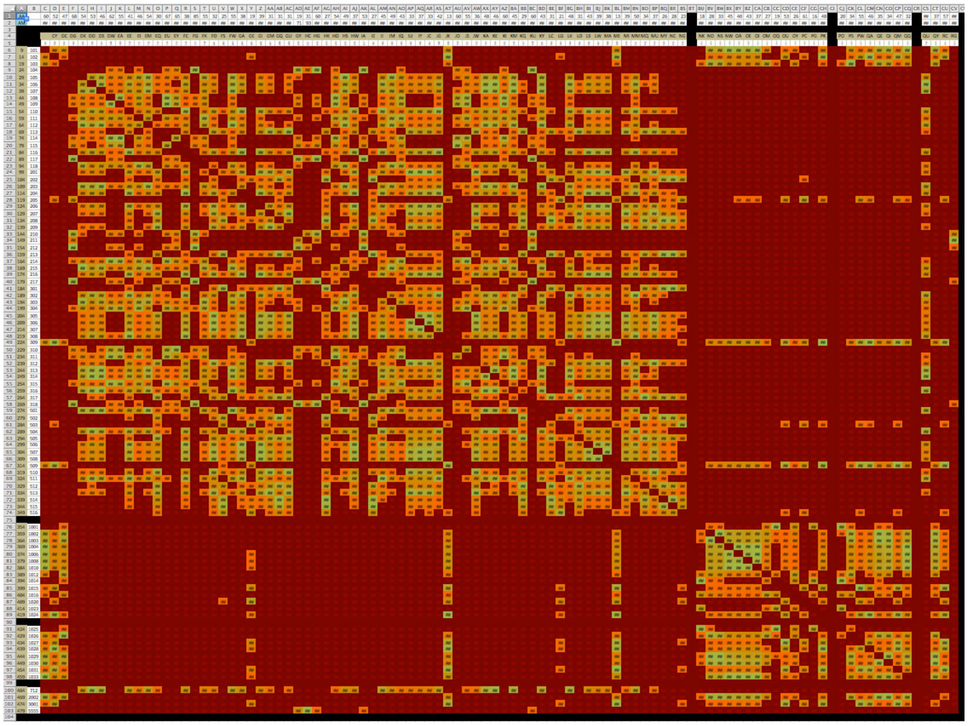


Figure 15: Heat map of imperfect Verbs

3.6. The Structure of Astérix

3.6.1. Distribution of Dialogue and Onomatopoeias

There is no substantial difference in the use of dialogue between the two authors, other than a difference in consistency. Uderzo used text less consistently on each page than Goscinny. The average standard deviation of word count per page for Uderzo is 59.7 (with an average of 168 words per page), while for Goscinny it is only 47.6 (with an average of 158 words per page).

The distribution of sound effects (text occurring outside of a speech bubble) was also relatively similar. The average number of onomatopoeias was similar (2.7 vs 2.5 per page), though Uderzo used them on five more pages per album on average (Goscinny using onomatopoeias on an average of 22.8 pages per album, Uderzo on 27.1). This contrast fits the standard deviation of words used per page, as those pages that feature more action tend to feature less plot development in the text, whether narration or dialogue.

3.6.2. Usage of Narration Boxes

Narration boxes, referred to as "captions" by Lyga and Lyga (Lyga and Lyga 161) are any textual elements that could be said to be non-diegetic; they are words intended solely for the reader, and no character within the *BD* in question is privy to them. Other forms of text in the *BD* are either visible or audible to at least one character (as in spoken text or onomatopoeic sound effects), and even thought bubbles are known to the character thinking. Narration boxes are most often used in *Astérix* to accomplish one of two goals: explaining things said, seen, or read within the strictly diegetic world of the *BD*, or to help bridge the transition between frames. In the case of the explanation, often the author will explain a historic concept, personage or deity, while in the case of the transition it is used to help make a scene-to-scene transition easier to follow (McCloud 71). In such transitions, the text will often help the reader understand where or when the scene is changing to; in this sense, from a purely statistical point of view, Goscinny prefers "*pendant ce temps*" while Uderzo seems more interested in using "*plus tard*."

The pacing of the albums written by each author can be summed up in this comparison. Goscinny used the expressions "*pendant ce temps*" and "*peu après*" 65% more often than Uderzo, and used the expression "*plus tard*" 48% less often than Uderzo. In terms of usage of narration boxes, Goscinny only used an average of 5 more than Uderzo per album, however, he also used more text in his narration boxes. In the end, this translates to 1/3 more text in narration boxes in Goscinny's work than in Uderzo's. This continues to paint a picture of the dichotomy of pages for Uderzo. It seems as though he vacillates between those pages that are more textual in nature that establish plot elements, and those pages which use fewer words and fewer narration boxes.

4. CONCLUSIONS

4.1. Nicolas and Astérix

The comparison of Goscinny's *Le petit Nicolas* and his *Astérix* brings interesting information to light. Though the comparison can't be expanded to a universal scope in comparing all traditional texts and *BD*, it does at the very least demonstrate potential differences between the media. What's more, it can serve as a demonstration that certain preconceptions about *BD* and comics *aren't* universally true.

The defining traits of *Astérix* (as compared to *Nicolas*) are: the usage of multiple punctuation, shorter sentences, fewer relative pronouns, a more lexically diverse inventory of words, a predilection for the present tense and the imperative, and slightly longer words. Starting with the easiest assessments, *Nicolas* uses more words per page than *Astérix*, though *Astérix* is the more lexically diverse work. Which means page for page, yes, *Astérix* can't compare to the amount of reading that one (presumably, an impressionable young child) would undertake. However, across the stretch of 100,000 words, reading *Astérix* will have exposed that impressionable young child to more vocabulary than *Nicolas*, which will have opted to re-use (and indeed, some might argue, *reinforce*) a smaller number of words. In short, the visual elements of the *BD* provide authors an economy of expression, allowing them to focus on the most important items to be written, leaving those plot point that are better left shown to the work of the artists.

The narrative style is completely different, as can be demonstrated through the comparison of verb tenses. Where *Astérix* tells a story directly, *Nicolas*'s stories are all told to the reader second-hand, most often from the point of view of the series' eponymous character who explains what has happened to him. Unlike many other kinds of literature, this *style raconté* found in *Nicolas* is in fact quite close to what happens in *BD*. Though stories can be told exclusively through pictures and narration, more often than not, they use the characters' conversations to allow the story to progress.

Interesting points of further analysis that won't be developed here also arise. Pedagogically, is one of these bodies better suited to a language learner's needs? At face value, it would certainly seem as though *Astérix* has its advantages, in that it has more graphic elements (allowing a learner to contextualize), it has a larger vocabulary, and a part of the storytelling itself is essentially non-verbal. However, that same vocabulary could hurt the learner, allowing them to focus far too often on words they might not see again, and ones that may have nothing to do with the day-to-day life and practical language skills that might be present in *Nicolas*.

4.2. Goscinny and Uderzo

Though it may seem rather obvious to critique Uderzo on the grounds that he was playing to his strengths, the composition of his albums of *Astérix* as compared to those of his collaborator Goscinny shows a less consistent approach. There are fewer diegetic references to earlier works through tried and true expressions. The variance in word count and the amount of text on each page is greater than it was when Goscinny was writing. This resulted in more pages that had less text. And though he included more words per album than his late partner, and perhaps even more complicated sentences (if the use of relative pronouns is any indication), it seems as though they were used less elegantly, and that their size and clout works against them as a heavier bit of narration that more forcefully gets his narrative from page to page. Indeed, the structure of Uderzo's work is both less consistent and "less coherent" as Norot assessed.

Also of note is the fact that his use of narration boxes points to a less tightly constructed narrative structure. Though some of his works, like *Astérix chez Rahàzade*, have a built-in time structure (in this case the looming sacrifice of the Indian princess), they still feel comparatively slower. Subjectively speaking, never has a crew of under-the-gun heroes seemed less pressed. By contrast, Goscinny's initial outing, *Astérix le Gaulois*, seems to obey the three classical unities rather well. The entire plot takes place in a day (save for the opening pages of establishing historical context), in or around the Gaulish village, and the entire plot revolves around the magic potion. Only in the last panel does the sun finally set, when the village is enjoying a feast. Again, as a subjective observation, even though Goscinny's other albums employ time differently, the pacing remains faster and the story doesn't seem to take as many sidesteps.

4.3. Comic Corpora

Without any doubt, the most difficult aspect in preparing a standard corpus of *BD* text is that the language itself is used more freely with regards to spelling conventions. Any sizeable corpus MUST be prepared by hand, containing both the original text as written and a glossed version of the text. This latter section would allow for analysis of thematic elements and subjects, though it could only be tenuously used to infer what words the author intended. In some cases, it is impossible to gloss what has been written, because the misusage of conventional language itself forms an inherent part of the communication. In Hergé's *Tintin au Congo*, for example, the natives use a specific grammatical construction, "y en a" to mean variously *il y a*, conjugations of the verb *être*, or indeed any number of other expressions to round out a phrase. In this sense, though the language input is what the author intended, it isn't conventional French, and won't fall into normal usage patterns.

Even when a selection of text can be glossed word for word with accepted forms, the use of nonstandard spelling for comedic effect, emphasis, or to build tension means that we can't simply reduce these words to their dictionary forms without losing meaning. Though the text can be separated from the image, in some cases, the text only exists as part of the visual field. The striking *baf!*, *bom!*, *tchoc!*, and *tchrâââk!* noises evoked by the words connect visual elements into their own sentences, a sort of one-word visual shorthand for such concepts as "Astérix pummels three Roman soldiers simultaneously" or "the rickety bridge begins to crumble underneath the feet of our heros." Any computerized study would have to inventory the meanings of each of these to do any sort of inventory of meaning and topos in a *BD*.

Beyond these concerns, the creation of any sort of glossed database for French *BD* would be prohibitive due to French copyright law. Though it would be interesting to compare works between authors, or indeed compare the work of an author across genres (such as Goscinny's *Lucky Luke* albums), the work is prohibitively time intensive for any one individual to undertake, and there are no fair use provisions which would allow for any such information to be shared so that another researcher might use or expand upon it.

4.4. Lemma Lists

The construction of a lemma list for use with *AntConc* posed several problems. Firstly, no lemma list existed in a format usable by the program. Acquiring a lemma list and formatting it took some effort. Secondly, the number of homographs in the French language posed a particular issue, as certain words like *être* and *suivre* have overlapping forms (in this case, identical forms for the first person singular present conjugation).

There are currently no tools that allow for the specific ordering of lemmas by frequency for those forms that are ambiguous. In this sense, *TreeTagger* could benefit from the implementation of a new part of speech tag that indicates that an item could be a form of a number of different lemmas. It could also stand to have improvements made to how it determines whether a word is a proper noun, as it seems overly dependent upon capitalization and only some entries were identifiable even without a capitalized first letter (it recognized “idéfix” as a proper noun yet not “astérix”).

4.5. Point of Departure

A number of subjects present themselves for additional study based on the initial results discussed here. The following sections briefly discuss those areas which could prove interesting or useful results should additional research be undertaken.

4.5.1. Potential Improvements to the TreeTagger Project

The accuracy results of how well TreeTagger was able to identify parts of speech are telling. The accuracy was far better for traditional texts than it was for *BD*. There seems to be a completely different style of text in *BD*, which is attempting to mimic the spoken language more closely. As such, an interesting hypothesis arises: would training the TreeTagger to more accurately recognize the transcriptions of spoken French benefit its ability to correctly parse *BD*. By extension, it would be interesting to see how well the text in various *BD* corresponds to texts in spoken corpora.

Additionally, the implementation of multiple parts of speech for each entry would be beneficial, as many entries are ambiguous. TreeTagger currently displays potential verbs (sommes being a conjugation of either *être* or *sommer*, for example) that it might get tripped up on, however, no such feature exists for parts of speech.

4.5.2. Pedagogical Applications

The comparative value of French language teaching using *Le petit Nicolas* or *Astérix* as supplemental readings could be approached in a number of ways. Firstly, a study of the levels of vocabulary contained in each section of the corpus for their practicality and utility could be undertaken by comparing the word lists generated to either a frequency dictionary or to the vocabulary presented in the textbook of a specific course of study. Secondly, a more traditional study of the kinds of phrases are in each section of the corpus could prove interesting. As mentioned above, the *BD* strives to emulate spoken language, not written language, and could prove a more interesting point of study for spoken language skills. Lastly, coupled with a more traditional study of *BD* structure, it would be interesting to catalog those visual elements that appear in the art as well as the text, serving as a potential point for reinforcement and demonstration of language use.

4.5.3. An Expanded Corpus

The addition of further *BD* series to the corpus might provide more interesting insights into the style of Goscinny's writings. He also wrote other series such as *Oumpah-pah, Iznogoud* and is credited as the *scénariste* of approximately 40 volumes of *Lucky Luke*. In particular, a study of N-grams would prove interesting, as there are periods of *Lucky Luke* both before and after Gocinny's tenure as *scénariste*, though he is credited with the introduction and subsequent re-use of such phrases and epithets as the one applied to Lucky Luke himself, "*l'homme qui tire plus vite que son ombre.*"

Expanding the corpus to incorporate other *BD* would be a massive undertaking at present, given the difficulty in digitizing the text, but it would be interesting to study the lexical density of texts in a range of *BD* works across different genres to see if the medium itself promotes elevated lexical density, or if *Astérix* only appears lexically dense when compared to *Le petit Nicolas*.

5. BIBLIOGRAPHY

ABBYY. *ABBYY FineReader*. Computer software. *OCR software for text recognition OCR PDF features*. Web. <http://finereader.abbyy.com/>.

Anthony, Laurence. *AntConc*. Computer Software. *Laurence Anthony's Software*. Web. (http://www.antlab.sci.waseda.ac.jp/software.html)

asterix.com. "Astérix Edition - La Grande Collection - Les Étapes De La Restauration - Le Site Officiel". Les Editions Albert-René/Goscinny-Uderzo. December 30, 2011 2011. <http://www.asterix.com/edition/la-grande-collection/etapes-restauration.html%3E.

---. "Astérix Encyclopédie - Les Traductions - Le Site Officiel". <http://www.asterix.com/encyclopedie/traductions/%3E.

---. "Astérix Galerie - Les Expositions - Les Archives D'albert Uderzo (Suite) - Le Site Officiel". January 11, 2012 2012. <http://asterix.com/galerie/expositions/archives-uderzo-suite.html%3E.

BD, MO'RE. "Top 50 Des Meilleures Ventes Bd 2008". 2009. December 31 2011 2011. <http://infobd.over-blog.com/article-27085295.html%3E.

---. "Top 50 Des Meilleures Ventes Bd En 2006". 2007. December 30, 2011 2011. <http://infobd.over-blog.com/article-5344933.html%3E.

Beuve-Méry, Alain. "La Bande Dessinée, Un Secteur En Bonne Santé." *Le Monde* 2009. Print.

Goscinny, René, and Albert Uderzo. *Astérix chez les Belges*. 1979. Reprint. Paris: DARGAUD, 1984. Print.

- - -. *Astérix chez les Bretons*. 1966. Reprint. Paris: DARGAUD, 1984. Print.

- - -. *Astérix chez les Helvètes*. 1970. Reprint. Paris: DARGAUD, 1984. Print.

- - -. *Astérix en Hispanie*. 1969. Reprint. Paris: DARGAUD, 1984. Print.

- - -. *Astérix et Cléopâtre*. 1965. Reprint. Paris: DARGAUD, 1984. Print.

- - -. *Astérix gladiateur*. 1964. Reprint. Paris: DARGAUD, 1984. Print.

- - -. *Astérix le Gaulois*. 1961. Reprint. Paris: DARGAUD, 1993. Print.

- - -. *La zizanie*. 1970. Reprint. Paris: DARGAUD, 1984. Print.

Goscinny, René, and Jean-Jacques Sempé. *Le petit Nicolas*. 1960. Reprint. Paris: Éditions Gallimard, 1991. Print.

- - -. *Le Petit Nicolas a des ennuis*. 1964. Reprint. Paris: Éditions Gallimard, 1994. Print.

- - -. *Les récrés du petit Nicolas*. 1963. Reprint. Paris: Éditions Gallimard Jeunesse, 2001. Print.

- - -. *Les vacances du petit Nicolas*. 1963. Reprint. Paris: Éditions Gallimard Jeunesse, 2001. Print.

Guillaume, Marie-Ange, and José-Louis Bocquet. *René Goscinny Biographie*. Actes Sud, 1997. Print.

Hanauer, David. *Word Counter*. Computer software. *Word Counter*. Web. <http://www.supermagnus.com/mac/Word_Counter/index.html>.

Imao, Yasu. *CasualTreeTagger*. Computer Software. *CasualTreeTagger – CasualConc*. Web. <https://sites.google.com/site/casualconc/utility-programs/casualtreetagger>.

Larson-Hall, Jenifer. *A Guide to Doing Statistics in Second Language Research Using SPSS*. Second Language Acquisition Research Series. 2 ed. New York: Routledge, 2010. Print.

Le cercle de la librairie. "Le Bulletin Du Livre." 1964. nos. 108-118 (1964): 25. Print.

Leloup, Roger. Yoko Tsuno l'intégrale, volume 1: De La Terre à Vinéa. Brussels: Dupuis, 2006. Print.

Lyga, Allyson A. W. and Barry Lyga. *Graphic Novels in Your Media Center: A Definitive Guide*. Libraries Unlimited, 2004. Print.

McCloud, Scott. *Understanding Comics: The Invisible Art*. Northampton, MA: Kitchen Sink Press, 1993. Print.

Nuance. *Dragon Dictate*. Computer software. *Nuance – Dragon Dictate for Mac*. Web. <http://www.nuance.com/for-individuals/by-product/dragon-for-mac/dragon-dictate/index.htm>.

The Cartoonist: Jeff Smith, Bone and the Changing Face of Comics. 2009. DVD July 21, 2009.

Ramage, Homer. "Good Reading for Bad." *Ottawa Citizen* April 13, 1956. Print.

Schmid, Helmut. *TreeTagger*. Computer Software. *TreeTagger*. Web. <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger>.

Schneidermann, Daniel. "Astérix Et Le Pouvoir Mediatix." *Libération* October 21, 2005 2005. Print.

Schofield, Hugh. "Should Asterix Hang up His Sword?" *BBC News* October 22, 2009 2009. Print.

Uderzo, Albert. *Astérix chez Rahàzade* . Paris: A. René, 1987. Print.

- - -. *Astérix et Latraviata* . Paris: A. René, 2001. Print.

- - -. *La galère d'Obélix* . Paris: A. René, 1996. Print.

- - -. *La rose et la glaive* . Paris: A. René, 1991. Print.

- - -. *Le ciel lui tombe sur la tête* . Paris: A. René, 2005. Print.

- - -. *Le fils d'Astérix* . Paris: A. René, 1983. Print.

- - -. *Le grand fossé* . Paris: A. René, 1980. Print.

- - -. *L'odyssée d'Astérix* . 1981. Reprint. Paris: A. René, 1984. Print.

UIMA. "Uima Lemmatizer Project". December 30, 2011. <http://code.google.com/p/uima-lemmatizer/%3E.

6. APPENDICES

Appendix A: Homographs Removed from UIMA Lemma List

Form(s)	Removed from Lemma	Added to Lemma
dans	dan	
as, es, ps, s	s	
c, ac	c	
ah, ch, eh, h	h	
cm, hm, m	m	
suis	suivre	
somme, sommes	sommer	
lui	luire	
maintenant	maintenir	
toute, toutes	touter	
monde	monder	
fois	foi	
arme, armes, armée, armées	armer	
heure, heures	heur	
galère, galères	galérer	
fous	foutre	
sous	sou	
pendant	pender	
tu	taire	
content	conter	
peuple, peuples	peupler	
table	tabler	
fils	fil	
poche, poches	pocher	
grave, graves	graver	
contente, contentes	contenter	
allions	allier	
tonne	tonner	
s		se
m		me
n		ne
t		te

The following lemmas were removed entirely due to either providing false positive matches (those words containing apostrophes), or being too easily confused with semantically-related tems.

contrer, barder, jusqu'au-boutiste, prud'homme, entr’aimer, patrouiller, huiler, imager, départir, peiner, s'agir, saler, casquer, blaguer, lamper, voiler, fauter, boucher, courser, pirater, baller, étoiler, miniaturer, miser, confiturer, baffer, gifler, pierrer, confiancer, vaguer, corser, parton

Appendix B: Word Count, Average Word and Sentence Length per Album

Text	Total Words	Avg Sentence Length	Avg Word Length	Relative Pronouns
Stories from *Le petit Nicolas*	1530	19.3671	3.9515	41
	1543	18.1529	3.7656	37
	1519	14.4667	3.7728	38
	1440	14.2574	3.9016	30
	1527	17.3523	3.857	29
	1667	22.527	3.6501	32
	1362	14.6452	3.9291	33
	1624	21.0909	3.7582	31
	1531	20.4133	3.8496	25
	1431	20.4429	3.6735	21
	1252	24.549	4.0281	32
	1623	18.4432	3.6583	22
	1442	18.9737	3.9355	27
	1555	14.8095	3.7908	41
	1545	21.4583	3.7796	33
	1338	23.069	3.6499	19
	1761	23.1711	3.9373	38
	1557	17.6932	3.7712	38
Stories from *Les récrés du petit Nicolas*	1244	15.55	4.0514	33
	1341	19.1571	3.5875	27
	1687	21.3544	3.6715	29
	1379	17.0247	3.8358	35
	1267	11.8411	3.6132	24
	1365	17.2785	3.7914	35
	1306	18.6571	3.8018	38
	1464	18.0741	3.6592	26
	1412	12.8364	3.852	35
	672	26.88	4.5952	17
	792	24	4.3526	14
	1229	19.8226	4	29
	1408	17.1707	3.7691	30
	1428	18.3077	3.8793	37
	1438	20.2535	3.8682	35
	1327	17.6933	3.9371	22
	1278	31.1707	3.9448	33

Text	Total Words	Avg Sentence Length	Avg Word Length	Relative Pronouns
Stories from *Les vacances du petit Nicolas*	1533	18.25	3.7694	34
	1513	16.8111	3.7887	29
	1447	20.0972	3.7425	29
	1310	17.2368	3.959	26
	1406	17.575	3.804	31
	1390	19.3056	3.8563	39
	1276	16.359	3.8182	27
	1349	17.2949	3.7168	25
	1442	14.42	3.7414	30
	1586	17.6222	4	43
	1571	16.7128	3.895	36
	1561	13.8142	3.6665	29
	1602	21.0789	3.8176	33
	1470	14	3.8263	42
	1538	15.8557	3.9004	27
	1488	14.1714	3.9119	27
	1459	17.369	3.8237	31
	1172	20.5614	3.9223	32
Stories from *Le petit Nicolas a des ennuis*	1422	18.96	3.7363	26
	1562	16.7957	3.8109	29
	1411	18.5658	3.6975	37
	1431	14.7526	3.8962	33
	1468	16.3111	4.0631	24
	1564	16.2917	3.8223	31
	1455	12.8761	3.8495	32
	1404	12.8807	3.7642	24
	1386	14	3.9598	31
	1626	17.2979	4.0962	40
	1746	17.46	3.762	25
	1727	15.844	3.8438	40
	1511	13.7364	3.8369	38
	1341	16.7625	3.9439	24
	1446	9.7703	3.8499	28
	1435	12.8125	3.8524	23
Le petit prince	15456	11.5602	4.072	257
Tintin au Congo	6740	6.7131	4.1987	97
Trio de l’étrange	7589	8.7532	4.5362	88
Le grand Meaulnes	67922	20.6137	4.4147	1459

Text	Total Words	Avg Sentence Length	Avg Word Length	Relative Pronouns
ASTERIX LE GAULOIS	5492	7.0773	4.2191	71
LA SERPE D'OR	5208	6.626	4.2696	95
ASTERIX ET LES GOTHS	6140	6.2717	4.4045	93
ASTERIX GLADIATEUR	7233	7.0773	4.2984	131
ASTERIX ET CLEOPATRE	6752	7.6727	4.3126	111
ASTERIX CHEZ LES BRETONS	7419	6.3519	4.3808	110
ASTERIX LEGIONNAIRE	6373	5.6902	4.2935	98
ASTERIX AUX JEUX OLYMPIQUES	6233	7.4026	4.3587	121
ASTERIX EN HISPANIE	6208	6.521	4.3063	92
LA ZIZANIE	6697	7.8236	4.2985	93
ASTERIX CHEZ LES HELVETES	6252	7.0564	4.3475	97
ASTERIX EN CORSE	6836	6.7549	4.3281	118
OBELIX ET COMPAGNIE	5487	7.794	4.1368	101
ASTERIX CHEZ LES BELGES	6165	6.9898	4.2383	106
LE GRAND FOSSE	6690	7.6985	4.3865	101
L'ODYSSEE D'ASTERIX	6952	7.38	4.3141	126
LE FILS D'ASTERIX	7010	6.9064	4.2703	141
ASTERIX CHEZ RAHAZADE	6718	7.185	4.2811	122
LA ROSE ET LE GLAIVE	6752	6.1104	4.4149	135
LA GALERE D'OBELIX	6299	7.0696	4.3469	111
ASTERIX ET LATRAVIATA	6289	6.5104	4.2217	120
LE CIEL LUI TOMBE SUR LA TETE	5182	6.7739	4.365	112

Appendix C: Word Count per Page (*Astérix* only)

Page Number	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Astérix le Gaulois	0	158	127	119	151	161	162	128	131	161	135	133	178	162	113
La Serpe d'or	0	92	181	160	153	133	112	117	102	74	145	131	163	82	51
Astérix et les Goths	0	180	108	144	140	205	189	201	159	133	174	128	185	131	140
Astérix Gladiateur	0	256	151	161	217	217	109	227	215	248	241	234	125	247	118
Astérix et Cléopâtre	0	196	227	213	239	224	232	227	166	189	146	132	131	79	142
Astérix chez les Bretons	0	163	327	264	184	212	223	196	162	197	224	183	194	123	112
Astérix Légionnaire	0	84	143	176	87	184	149	123	153	202	169	117	139	163	150
Astérix aux jeux Olympiques	0	187	181	250	75	196	150	186	181	114	178	155	159	182	84
Astérix en Hispanie	0	205	154	94	217	191	156	66	186	127	235	170	132	135	81
La Zizanie	0	190	298	139	181	258	104	149	232	172	118	93	147	168	206
Astérix chez les Helvètes	0	109	127	174	190	136	194	173	232	220	200	187	252	239	110
Astérix en Corse	241	74	137	178	182	177	250	144	178	133	143	296	189	78	222
Obélix et Compagnie	0	67	101	72	131	91	116	92	198	161	139	193	100	125	112
Astérix chez les Belges	0	149	168	157	126	230	240	172	104	139	159	45	151	106	79
Le Grand Fossé	0	99	91	303	140	79	166	157	52	353	130	261	112	208	298
L'Odyssée d'Astérix	0	192	129	191	245	166	159	73	146	271	206	108	103	150	263
Le Fils d'Astérix	0	167	116	192	147	158	155	302	151	164	124	241	171	207	219
Astérix chez Rahazade	0	257	153	169	148	174	317	146	127	186	312	171	150	142	182
La Rose et le glaive	0	140	183	216	283	197	279	203	136	187	220	239	164	197	177
La Galère d'Obélix	0	86	196	136	178	195	157	123	114	125	122	143	226	106	156
Astérix et Latraviata	0	71	113	196	176	256	193	227	153	204	183	149	111	169	204
Le Ciel lui tombe sur la tête	0	189	102	244	147	13	148	125	157	160	64	140	126	305	168

Page Number	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**
Astérix le Gaulois	80	115	122	102	130	148	162	105	135	127	114	128	117	196	171
La Serpe d'or	157	227	207	170	118	147	75	120	128	208	188	168	117	124	107
Astérix et les Goths	149	200	171	178	196	156	125	150	166	101	142	131	190	144	161
Astérix Gladiateur	133	228	160	198	248	175	120	179	127	124	202	167	143	147	162
Astérix et Cléopâtre	214	201	128	160	167	168	197	145	176	120	176	128	216	232	145
Astérix chez les Bretons	194	200	151	166	198	173	174	181	149	147	210	197	133	202	169
Astérix Légionnaire	154	162	166	156	117	161	175	136	184	165	165	174	128	111	160
Astérix aux jeux Olympiques	185	185	105	153	104	137	150	236	166	159	154	134	151	179	120
Astérix en Hispanie	90	192	202	218	243	138	230	115	120	230	75	189	128	97	166
La Zizanie	292	148	156	83	169	86	210	197	87	195	179	127	134	140	138
Astérix chez les Helvètes	198	145	180	139	166	229	152	217	154	79	192	167	125	220	198
Astérix en Corse	98	111	157	207	79	156	167	266	175	212	168	190	79	210	209
Obélix et Compagnie	191	142	147	122	143	187	104	91	122	203	157	211	70	130	162
Astérix chez les Belges	208	111	166	171	185	134	99	159	199	208	176	186	193	175	125
Le Grand Fossé	286	157	220	212	197	170	165	328	125	184	123	188	142	109	238
L'Odyssée d'Astérix	264	214	187	246	169	163	174	162	138	253	215	176	281	97	162
Le Fils d'Astérix	267	128	196	215	217	169	316	222	238	40	205	271	159	262	148
Astérix chez Rahazade	174	149	165	117	234	138	200	235	275	130	143	239	110	204	132
La Rose et le glaive	128	204	167	181	197	288	176	254	25	92	189	181	140	186	141
La Galère d'Obélix	196	172	93	201	231	208	139	131	136	205	239	162	194	136	294
Astérix et Latraviata	157	238	181	190	123	196	187	120	112	141	80	92	70	301	176
Le Ciel lui tombe sur la tête	69	138	192	94	66	212	35	103	62	93	189	225	156	205	31

Page Number	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**
Astérix le Gaulois	181	128	121	62	80	93	109	141	195	84	121	81	54	126	172
La Serpe d'or	104	111	203	127	146	142	77	110	91	149	152	210	221	0	0
Astérix et les Goths	148	153	136	185	111	67	218	134	131	134	178	532	111	179	0
Astérix Gladiateur	225	237	257	201	155	162	146	184	164	128	157	137	214	173	106
Astérix et Cléopˆâtre	147	220	127	222	181	141	141	203	125	98	205	150	154	236	105
Astérix chez les Bretons	154	202	317	165	84	54	129	111	96	136	268	262	254	226	253
Astérix Légionnaire	222	163	168	159	206	206	249	244	197	173	208	150	173	157	88
Astérix aux jeux Olympiques	124	155	172	141	203	132	165	164	220	151	165	169	107	140	256
Astérix en Hispanie	164	114	100	163	230	121	206	185	92	190	310	97	72	187	199
La Zizanie	225	85	179	162	206	172	148	90	335	128	174	200	157	161	170
Astérix chez les Helvètes	114	141	159	108	142	148	169	215	99	83	50	131	77	118	120
Astérix en Corse	194	199	130	183	143	193	150	173	99	249	150	150	194	149	123
Obélix et Compagnie	181	215	227	125	115	137	206	194	127	152	110	133	107	63	123
Astérix chez les Belges	149	138	172	169	144	169	98	184	128	150	36	157	175	10	96
Le Grand Fossé	213	178	139	135	17	52	142	226	203	142	210	156	152	145	162
L'Odyssée d'Astérix	186	235	120	116	106	157	181	213	157	227	198	214	96	269	171
Le Fils d'Astérix	190	235	195	144	121	111	184	51	103	116	200	148	195	190	230
Astérix chez Rahazade	288	175	231	258	162	158	66	131	131	206	111	99	106	91	222
La Rose et le glaive	138	211	192	127	208	127	110	73	195	158	93	205	103	122	245
La Galère d'Obélix	161	256	139	153	200	296	135	106	135	68	216	91	161	64	209
Astérix et Latraviata	167	214	222	80	84	130	156	160	232	149	131	69	193	188	243
Le Ciel lui tombe sur la tête	99	10	97	125	164	218	156	80	182	121	118	172	50	44	250

Appendices D - F: Part of Speech Counts, Pairwise Chi-Square Analyses of Part of Speech Usage, and Electronic Access to Data

The full range of data generated is too cumbersome to be formatted into reasonable tables. As such, the full range of data can be acquired electronically upon request. Please contact Dennis Meyer at dennis.s.meyer@gmail.com with a specific request for the data required, as well as a short explanation of your interest in the data. Due to copyright concerns, certain items cannot be furnished (for example, texts still protected by copyright), though every reasonable effort will be made to provide data that will be used for academic purposes.

Appendix G: Lexical Diversity

Series	Title	Number of Unique Lemmas	Number of Words
Astérix by Goscinny	Astérix le gaulois	1034	5434
	Astérix et la serpe d'or	984	5131
	Astérix chez les Goths	1175	6079
	Astérix gladiateur	1256	7148
	Astérix et Cléopâtre	1187	6729
	Astérix chez le Bretons	1294	7337
	Astérix légionnaire	1136	6297
	Astérix aux jeux Olympiques	1188	6180
	Astérix en Hispanie	1143	6152
	La Zizanie	1146	6621
	Astérix chez les Helvètes	1171	6191
	Astérix en Corse	1187	6802
	Obélix et Compagnie	933	5444
	Astérix chez les Belges	1079	6128
Astérix by Uderzo	Astérix et le Grand Fossé	1272	6643
	L'Odyssée d'Astérix	1357	6889
	Le Fils d'Astérix	1251	6923
	Astérix chez Rahàzade	1333	6665
	La rose et le glaive	1313	6692
	La galère d'Obélix	1217	6235
	Astérix et Latraviata	1165	6227
	Le ciel lui tombe sur la tête	1053	5130
Le petit Nicolas	Un souvenir qu'on va chérir	324	1529
	Les Cow-boys	328	1542
	Le Bouillon	308	1517
	Le football	306	1442
	On a eu l'inspecteur	317	1525
	Rex	309	1661
	Djodjo	321	1358
	Le chouette bouquet	315	1619
	Les carnets	339	1528
	Louisette	293	1429
	On a répété pour le ministre	306	1251
	Je fume	316	1621
	Le petit poucet	331	1442
	Le vélo	331	1552
	Je suis malade	336	1544
	On a bien rigolé	282	1337
	Je fréquente Agnan	380	1761
	M. Bordenave n'aime pas le soleil	313	1552

Series	Title	Number of Unique Lemmas	Number of Words
Les récrés du petit Nicolas	Alceste a été renvoyé	319	1242
	Le nez de tonton Eugène	265	1333
	La montre	343	1684
	On fait un journal	302	1378
	Le vase rose du salon	283	1260
	À la récré on se bat	299	1365
	King	293	1304
	L'appareil de photo	315	1459
	Le football	298	1413
	1re mi-temps	436	1465
	2e mi-temps		
	Le musée de peintures	293	1228
	Le défilé	334	1401
	Les boy-scouts	309	1429
	Le bras de Clotaire	315	1436
	On a fait un test	321	1327
	La distribution des prix	305	1280
Les vacances du petit Nicolas	C'est papa qui décide	339	1534
	La plage c'est chouette	336	1508
	Le boute en train	306	1446
	L'île des Embruns	326	1310
	La gym	300	1404
	Le golf miniature	277	1389
	Le "on a joué à la marchande"	269	1271
	On est rentrés	270	1344
	Il faut être raisonnable	313	1440
	Le départ	363	1584
	Courage	343	1574
	La baignade	311	1561
	La pointe des Bourrasques	346	1601
	La sieste	296	1466
	Jeu de nuit	349	1536
	La soupe de poisson	327	1489
	Crépin a des visites	322	1453
	Souvenirs de vacances	265	1171
Le petit Nicolas est malade	Joachim a des ennuis	303	1422
	La lettre	311	1555
	La valeur de l'argent	286	1402
	On a fait le marché avec papa	312	1429
	Les chaises	323	1468
	La lampe de poche	322	1558

Series	Title	Number of Unique Lemmas	Number of Words
Le petit Nicolas est malade	La roulette	292	1454
	La visite de mémé	322	1398
	Leçon de code	309	1370
	Leçon de choses	350	1629
	À la bonne franquette	338	1743
	La tombola	329	1723
	L'insigne	302	1510
	Le message secret	321	1339
	Jonas	295	1446
	La craie	313	1427
Le Grand Meaulnes		4695	72286
Le petit prince		1737	16409
Tintin au Congo		1254	6877
Le trio de l'étrange		1743	8099