World Journal of Social Science Research
ISSN 2375-9747 (Print) ISSN 2332-5534 (Online)
Vol. 4, No. 2, 2017
www.scholink.org/ojs/index.php/wjssr

A Corpus-Based Analysis of the Language Used by Defendants of Homicide in Court

Katerina T. Frantzi[1*] & Anastasia K. Katranidou[1]

[1] Department of Mediterranean Studies, University of the Aegean, Rhodes, Greece

[*] Katerina T. Frantzi, E-mail: frantzi@rhodes.aegean.gr

Received: April 10, 2017 Accepted: May 5, 2017 Online Published: June 6, 2017

doi:10.22158/wjssr.v4n2p164 URL: http://dx.doi.org/10.22158/wjssr.v4n2p164

Abstract

In this study we present the updated version of the Greek Corpus of Defendants' Testimonies, GCDT and a series of new evaluations that have been carried out on the defendants' speech. Using criteria, such as lexical richness, lexical density, part-of-speech frequencies, word and sentence length, we look for linguistic features which could characterize the stylometric profile of the defendants. We also present GCWT, a reference corpus that has been constructed similar to GCWT stylistic features. GCWT contains witnesses' testimonies collected in the court.

Keywords

forensic linguistics, corpus-based studies, greek court, defendants' testimonies

1. Introduction

Forensic linguistics is the analysis of the language that relates to law, either as evidence or as legal discourse (Olsson & Luchjenbroers, 2013). Language as legal discourse includes, among others, the discourse inside the court room. The legal language can be divided in the professional language of law and the language of law encountered by the lay person (Gibbons, 2003).

Crime profiling is the identification of specific characteristics of an individual committing a crime by a thorough systematic observational process and an analysis of the crime scene, the victim, the forensic evidence and the known facts of the crime. The profiling technique is used by behavioral scientists and criminologists to examine criminal behavior and to evaluate or even predict future criminal actions (Davis, 1996). Author profiling or characterization is the procedure of extracting information about the age, education, sex, etc. of the author of a given text (Koppel et al., 2002). By combining these two objectives, we attempt to represent the general properties of the criminal's language style. Recent approaches for authorship attribution and author profiling have been examined by Stamatatos (Stamatatos, 2009), who evaluated characteristics for both text representation and text classification focusing on computational requirements. The research of Broussalis, Markopoulos and Mikros, which

presented the most distinctive stylometric characteristics, concluded that legal texts have a distinct and highly recognizable stylometric profile (Broussalis et al., 2012).

Katranidou and Frantzi (2016) introduced GCDT, the first Forensic Linguistics Greek Corpus that consists of defendants' testimonies collected in a natural environment. The first processing of GCDT showed that, compared to the general language, the defendants use some unusual words in their testimonies.

In this study, we first present the updated version of GCDT, the Greek Corpus of Defendants' Testimonies and GCWT, a reference corpus with similar stylistic features with GCDT. GCWT contains witnesses' testimonies collected in the court and related to homicide cases. We then present a statistical analysis based on standard stylometric features of the language used by defendants of homicide, inside the Greek court room, derived from the exchanges of discourse between them and the prosecutors.

2. The Updated Version of GCDT

GCDT is the first Greek Corpus of Defendants' Testimonies. The defendants have been accused of felony. The processing of the corpus showed that the defendants use some unusual words in their testimonies. Comparing the frequencies of the most frequent nouns and verbs to those of a general Greek corpus, it was found that defendants quite frequently use specific nouns and verbs which are rare in general language (Katranidou & Frantzi, 2016).

Our first goal was to extend GCDT with additional testimonies of similar criminal cases for the improvement of our statistical results. Thus, we updated the corpus with new testimonies which were gathered from the Court of Law of the Greek city of Thessaloniki. The updated version of GCDT consists of 109,523 words from 86 hearings, issued by 124 subjects, all of which are defendants of homicide. One hundred and ten of them are men and fourteen are women. Ninety-one are native Greek speakers and thirty-three testify through an interpreter (Note 1). Their average age at the time of the hearing is approximately 38 years. Their level of education is not precisely known. Regarding their occupation, most of them are workers, farmers, builders, freelancers, two are students, four are pensioners and twenty-four are unemployed.

In most of the cases (88.8%) the verdict is condemnatory and only in a few cases (11.2%) has the defendant been acquitted. The acquittals in homicides are much rarer than the convictions since the defendant's lawyer usually tries to find extenuating circumstances to reduce the defendant's penalty instead of aiming for an acquittal. The few times that the verdict is not condemnatory are due to lack of clear evidence for the crime.

3. Stylometric Features

3.1 GCDT and Hellenic National Corpus

We first repeated the usual statistical measurements, i.e. we focused on some part-of-speech frequencies, such as nouns and verbs, as well as on the most frequently used words including function words. We used Word Smith Tools v.5 (Scott, 1998) for processing the corpora.

Comparison of frequencies of GCDT with those of a reference corpus gives us information regarding special characteristics of the testimonies' language. We used the Relative Frequency measure (*fi*) since it provides the normalized frequency of every word in the corpus. By using relative instead of absolute frequency, we are given the capability of comparisons between the frequency of specific words in GCDT and a reference corpus. We examined the twenty most frequent nouns and verbs of GCDT and we compared their relative frequency with a Greek general language corpus. As a reference corpus we used the Hellenic National Corpus HNC (Note 2) (Hatzigeorgiu et al., 2000). HNC is currently the biggest written corpus of Modern Greek and consists of 50,824 texts and 47,013,924 words derived from written language material, such as books, newspapers, journals etc. The results showed a large variance in the frequencies of occurrence of words between the two corpora, not only for nouns where we would expect a higher frequency of occurrence in GCDT for specific words such as 'knife', 'money', 'gun', 'police', 'prison', but for other nouns used such as 'telephone' and 'mother'. Apart from the noun 'years', none of the other nineteen most common nouns in GCDT is as common in HNC but, on the contrary, they present a much lower frequency of occurrence.

Similarly, apart from the verb 'to be' ('is/are'), which is significantly less frequent compared to HNC corpus, the rest of the most frequently used verbs in GCDT, in present and past tense, are much rarer in HNC. It is worth noting that among the twenty most frequent verbs, fifteen are used in the past tense, since the defendants' testimonies describe a past action, i.e. 'was/were', 'I said', 'he/she said' and only five of them are used in present tense, relating to the hearing procedure: 'is/are', 'I know', 'I have', 'I remember', 'I am'.

Following verbs and nouns, we extracted the frequency lists of adjectives, adverbs and pronouns, and we compared their frequencies of occurrence to those in HNC. Regarding the use of adjectives, we noticed that, among the ten most frequent, there are simple adjectives such as 'first', 'second', 'third', 'many', 'good', 'small', 'big' and 'sure', which are quite common in HNC as well, and adjectives such as 'beaten' and 'drunk' which are considerably infrequent in HNC.

Adverbs seem to be used more frequently in GCDT than in HNC, since the defendants' language tends to be descriptive. The adverbs 'after', 'when', 'there', 'together', 'up', 'in', 'nice', 'before', 'much' and 'out' are the ten most frequently used. Apart from the adverb 'much', the rest present a much higher frequency of appearance compared to that of HNC.

Regarding pronouns, the two most frequent ones in GCDT, 'my' and 'I', are a lot rarer in HNC.

However, the pronouns ‘his’, ‘where’ and ‘her’, have much lower frequencies compared to those in HNC.

3.2 GCDT and GCWT

The first reference corpus, HNC, as mentioned above, consists exclusively of written language material and aims to be representative of the Greek general language. However, the defendants use specific vocabulary during the trial procedure. To achieve more accurate statistical results and be methodologically correct, we constructed a reference corpus with similar stylometric features to our study corpus, GCDT. The new reference corpus, GCWT (Greek Corpus of Witnesses’ Testimonies), consists of 395,925 words and has derived from witnesses’ testimonies related to homicide cases. Both GCDT and GCWT have been constructed from the transcriptions of the court spoken language during the trial procedure. The size of the reference corpus is four times greater than the study corpus, quite close to the ideal size of a reference corpus (Berber-Sardinha, 2000; Koppel et al., 2002).

4. Word List Derived Analyses

In order to define the stylometric profile of the GCDT and GCWT, we firstly measured some sets of stylometric features which are based on word list derived analyses. The features used in this study are the following:

4.1 Most Frequent Words

The WordSmith WordList tool gave us a list of all the words in GCDT in frequency order. As we expected, the top of this list is occupied by function words, such as ‘and’, ‘the’, ‘to’, ‘not’, ‘with’, ‘that’ etc., with the word ‘and’ holding the 4% of the total corpus size (Table 1). The most frequent 15 words in the list take up approximately one third of the corpus.

Table 1. Most Frequent Words in GCDT

s/n	word	freq.	freq. %	cumulative freq. %
1	and	4449	4,06	4,06
2	the	4031	3,68	7,74
3	to	3709	3,39	11,13
4	not	3612	3,30	14,43
5	me	2973	2,71	17,14
6	with	2468	2,25	19,40
7	him	2101	1,92	21,31
8	her	1858	1,70	23,01
9	that	1644	1,50	24,51
10	into	1608	1,47	25,98

11	he	1568	1,43	27,41
12	these	1524	1,39	28,80
13	I	1514	1,38	30,18
14	him	1438	1,31	31,50
15	for	1349	1,23	32,73

4.2 Lexical Richness

The lexical richness of a text accounts for how many different word types are used in the text. Table 2 shows the percentage of word types with frequency one and two in the corpus, namely the hapax and dis legomena and the ratio of dis legomena to hapax legomena in the text segment, which is indicative of the authorship style (Hoover, 2003). It is depicted that hapax legomena seem to take up almost 50% of the word types.

Table 2. Lexical Richness of GCDT and GCWT

	Hapax Legomena %	Dis legomena %	Dis-/Hapax- legomena
GCDT	4.61	15.4	0.31
GCWT	45.96	15.47	0.34

4.3 Part of Speech Frequencies and Lexical Density

Style is also characterized from the Part-of-Speech (POS) frequencies (Gamon, 2004; Zhao & Zobel, 2005). For this purpose, we used a Greek POS tagger (Note 3) and we measured the relative frequencies of content words (nouns, verbs, adjectives and adverbs) as well as function words (pronouns, articles, prepositions etc.). Lexical density, evaluating the proportion of content words in the text, is a measure of how informative a text is (García & Martin, 2007). For instance, spoken texts tend to have a lower lexical density (near 45%) than written ones (above 50%) (Johansson, 2008; Fan & Thomas, 2013; Ure, 1971). The content words' frequencies, function words' frequencies and lexical density of GCDT and GCWT are given in Table 3.

Table 3. Content Words' Frequencies, Function Words' Frequencies and Lexical Density of GCDT and GCWT

	content words' frequency %	function words' frequency %	Lexical Density %
GCDT	44.21	55.7	44.2
GCWT	45.83	54.1	45.8

Both GCDT and GCWT have low lexical density compared to the typical lexical density of written texts since they result from transcriptions of spoken language and are made of special language

material. The reference corpus has higher lexical density than GCDT, justified from the fact that GCWT contains testimonies from specialized witnesses, such as forensic pathologists and police officers, who tend to use more descriptive language and more information-bearing content words.

4.4 Word and Sentence Length and Standard Deviation

Standard deviations of both word and sentence length can also give information on how the defendants use language. Having made the appropriate measurements, we found that there are slight differences between the defendants' and the witnesses' speech as shown in Table 4.

Table 4. Word and Sentence Length and Standard Deviation of GCDT and GCWT

	Average word length in letters	Word length standard deviation	Average sentence length in words	Standard deviation of sentence length
GCDT	4.44	2.27	8.27	6.32
GCWT	4.64	2.54	8.76	6.46

There is a small difference in word length, yet witnesses seem to use more and larger words more frequently than the defendants. The average sentence length for defendants is shorter than that of witnesses, as is the standard deviation (6.32 words) for defendants compared to witnesses (6.46 words). Considering the nature of both corpora, the low standard deviations are not surprising. Both corpora have derived from testimonies inside a court and apart from some descriptive speech pieces, they contain responses. Typically, defendants and witnesses use one-word or short responses. Moreover, the defendants' educational level average is lower than the witnesses' and thus they tend to use simpler words and shorter sentences.

5. Keywords Derived Analyses

In order to perform a more qualitative content analysis, we used an approach based on keywords derived analyses. We used the WordSmith KeyWords tool to compare the word list extracted from our study corpus, GCDT, to a word list extracted from a reference corpus (Scott, 2001). The result of this comparison is the *keyness* value, which describes the value of a word being a 'key' in its context. Keywords are "items of unusual frequency in comparison with a reference corpus" (Scott & Tribble, 2006).

Table 5. List of the Keywords with Maximum Positive Keyness of GCDT

N	Keyword	Translation	Freq.	%	RC. Freq.	RC. %	Keyness
1	Είπα	I said	1072	0.98	1152	0.29	645.334
2	Πήρα	I took	428	0.39	218	0.05	472.894
3	Είχα	I had	701	0.64	829	0.20	443.144
4	έκανα	I did	340	0.31	161	0.04	421.269
5	Πήγα	I went	596	0.54	659	0.16	407.386
6	Ήθελα	I wanted	266	0.24	114	0.02	405.019
7	Να	To	3709	3.39	8799	2.22	396.663
8	Εγώ	I	1514	1.38	3040	0.76	315.245
9	κτύπησα	I hit	132	0.12	31	-	246.743
10	Μου	My	2973	2.71	6748	1.70	222.603
11	Πάω	I go	215	0.20	156	0.04	213.337
12	Κάνω	I do	164	0.15	105	0.03	175.852
13	μπορούσα	I could	117	0.11	67	0.02	147.743
14	ήμουν	I was	433	0.40	507	0.13	146.711
15	έφυγα	I left	139	0.13	102	0.03	131.471
16	φοβήθηκα	I got scared	87	0.08	38	-	129.232
17	έβαλα	I put	89	0.08	31	-	128.372
18	Θα	Will	892	0.81	1716	0.43	124.944
19	έπαιρνα	I was taking	93	0.08	49	0.01	114.531
20	Με	With	2468	2.25	6203	1.56	113.130
21	σκέφθηκα	I thought	56	0.05	15	-	105.390
22	σκοτώσω	Kill	65	0.06	18	-	94.969
23	Πάμε	we go	140	0.13	131	0.03	94.926
24	Ναι	Yes	322	0.29	503	0.13	91.392
25	πήγαινα	I was going	107	0.10	77	0.02	89.390

Table 5 depicts the list of the first 25 keywords with maximum positive keyness. The field ‘Freq’ is the frequency of the keyword in the study corpus, ‘%’ is its relative frequency in the study corpus, ‘RC Freq’ is the frequency of the keyword in the reference corpus, ‘RC %’ is the relative frequency in the reference corpus and ‘keyness’ stands for the value of the log-likelihood statistics.

These keywords are unusually frequent compared to GCWT and mainly consist of verbs in the first person, singular number, past tense. They are used to describe an action or a feeling. Table 6 depicts the 25 keywords with the greater negative value of keyness, i.e. keywords quite infrequent compared to the

reference corpus.

Table 6. List of Keywords with the Maximum Negative Keyness of GCDT

N	Keyword	Translation	Freq.	%	RC Freq.	RC. %	Keyness
1	Είχε	he/she had	490	0.45	3226	0.81	-323.641
2	κατηγορούμενος	defendant (he)	70	0.06	1279	0.32	-289.073
3	Ήταν	was/were	1338	1.22	5955	1.50	-262.056
4	Μας	Us	247	0.23	2138	0.54	-215.345
5	Ο	the (he)	1568	1.43	8002	2.02	-176.180
6	Η	the (she)	850	0.78	4817	1.21	-170.918
7	κατηγορούμενο	the defendant	34	0.03	644	0.16	-148.774
8	Ότι	That	1644	1.50	7869	1.98	-146.845
9	Του	His	1438	1.31	6682	1.68	-115.631
10	Θύμα	Victim	72	0.07	772	0.19	-105.168
11	κατηγορουμένου	defendant's	7	0.01	283	0.07	-94.113
12	κατηγορουμένη	defendant (she)	2	0.00	266	0.06	-92.636
13	Γνωρίζω	I know	34	0.03	422	0.11	-89.828
14	Γιος	Son	21	0.02	369	0.09	-86.989
15	Είχαν	they had	75	0.07	574	0.14	-74.823
16	Βρέθηκε	was found	11	0.01	233	0.06	-69.817
17	Είναι	is/are	612	0.56	2600	0.66	-66.381
18	Της	Her	512	0.47	2555	0.64	-64.358
19	Από	From	1167	1.07	4694	1.18	-60.167
20	Βρήκαμε	we found	9	0.01	164	0.04	-58.678
21	Υπηρεσία	Duty	1	0.00	154	0.03	-58.602
22	Οι	the (they)	239	0.22	427	0.11	-55.799
23	Άκουσα	I heard	54	0.05	491	0.12	-53.412
24	Των	Their	29	0.03	335	0.08	-50.940
25	Έκανε	he/she did	155	0.14	914	0.23	-50.318

The two keywords 'defendant' and 'victim' appear quite often in the reference corpus compared to GCDT, since defendants rarely refer to these two terms.

6. Conclusions and Future Work

In this work, we presented the updated version of GCDT that now includes a large number of

testimonies. At first, we carried out evaluations on the defendants' speech and found that defendants use generally infrequent words quite frequently, mostly nouns relative to crimes. Verbs are mainly found in past tense and adverbs are used quite often since the defendants' language tends to be descriptive. Then, we introduced GCWT, a reference corpus consisting of witnesses' testimonies in order to be able to make comparisons between the two corpora regarding stylometric features. Regarding the word list derived analysis, we noticed that defendants and witnesses have low lexical density compared to the typical lexical density of written texts. The reference corpus seems to be denser and the explanation could derive from the fact that it contains testimonies from specialized witnesses, who tend to use richer language. Furthermore, a keyword list derived analysis showed that some keywords of the GCDT corpus are unusually frequent compared to the GCWT corpus, even if both corpora have similar stylistic features.

Regarding future work, this includes the updating of the corpus with testimonies of the defendants during the criminal investigation. The updating of the corpus would allow us to compare the language used by defendants inside and outside the court. The second component of future work involves the exploration of features borrowed from the research areas of Information Retrieval and Language Modeling (Houvardas & Stamatatos, 2006; Mikros, 2012).

References

Berber-Sardinha, T. (2000). Comparing corpora with WordSmith Tools: How large must the reference corpus be? In *Proceedings of the Workshop on Comparing Corpora* (Vol. 9, pp. 7-13). Hong Kong: Association for Computational Linguistics.

Broussalis, G., Markopoulos, G., & Mikros, G. (2012). Stylometric profiling of the Greek Legal Corpus. *Selected Papers of the 10th International Conference of Greek Linguistics*, *10*, 167-176.

Davis, J. A. (1996). *Crime scene investigative analysis: Elements of profiling*. San Diego, CA: Author.

Gamon, M. (2004). Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In *Proceedings of the 20th International Conference on Computational Linguistics* (pp. 611-617).

García, A. M., & Martin, J. C. (2007). Function words in authorship attribution studies. *Literary and Linguistic Computing*， *22*(1), 49-66.

Gibbons, J. (2003). *Forensic Linguistics. An Introduction to Language in the Judicial System.* Oxford: Blackwell.

Hatzigeorgiu, N., Gavrilidou, M., Piperdis, S., Carayannis, G., Papakostopoulou, A., Athanasia, S., & Iason, D. (2000). Design and implementation of the online ILSP Greek Corpus. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperdis, & G. Stainhaouer (Eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation* (Vol. 3, pp. 1737-1742). Athens,

Greece: ELRA.

Hoover, D. (2003). Another perspective on vocabulary richness. *Computers and the Humanities, 37*, 151-178.

Houvardas, J. & Stamatatos, E. (2006). N-Gram Feature Selection for Authorship Identification. In Euzenat, J., & Domingue, J (Eds.), *Artificial Intelligence: Methodology, Systems, and Applications*, (Vol. 4183, pp. 77-86). Berlin/Heidelberg: Springer.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working Papers, 53*, 61-79.

Katranidou, A., & Frantzi, K. (2016). The Greek Corpus of Defendants'Testimonies: Frequent use of infrequent words. *European Journal of Humanities and Social Sciences* 3, 25-29.

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing, 17*(4), 401-412.

Mikros, G. K. (2012). Authorship Attribution and Gender Identification in Greek Blogs. *Methods and Applications of Quantitative Linguistics*, *21*, 21-32.

Olsson, J., & Luchjenbroers, J. (2013). *Forensic linguistics*. London: A&C Black.

Scott, M. (1998). *WordSmith Tools Version* 3. Oxford: Oxford University Press.

Scott, M. (2001). Comparing corpora and identifying key words, collocations and frequency distributions through the WordSmith Tools suite of computer programs. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small Corpus Studies and ELT* (pp. 47-67). Amsterdam/Philadelphia: John Benjamins Publishing Co.

Scott, M., & Tribble, C. (2006). *Textual Patterns: Keyword and corpus analysis in language education*. Amsterdam: Benjamins.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology, 60*(3), 538-556.

To, V., Fan, S., & Thomas, D. P. (2013). Lexical density and Readability: A case study of English Textbooks. *The International Journal of Language, Society and Culture*, *37*(7), 61-71.

Ure, J. (1971). Lexical density and register differentiation. In G. Perren, & J. L. M. Trim (Eds.), *Applications of Linguistics*, (pp. 443-452). London: Cambridge University Press.

Zhao, Y., & Zobel, J. (2005). Effective and scalable authorship attribution using function words. *Proceedings of the 2nd Asia Information Retrieval Symposium*. AIRS.

Notes

Note 1. There is an unavoidable loss of precision on the defendant's speech during the transcription procedure. In the case where an interpreter in used, the loss in precision on the defendant's speech is even bigger.

Note 2. Hellenic National Corpus, Institute for Language and Speech Processing, ATHENA Research & Innovation Information Technology, http://hnc.ilsp.gr.

Note 3. Natural Language Processing Group, Department of Informatics—Athens University of Economics and Business, http://nlp.cs.aueb.gr/software.html.