International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 16 No. 1 Jun. 2015, pp. 252-259

http://www.ijisr.issr-journals.org/

Performance Evaluation of Various Feature Extraction and Classification Techniques for Authorship Attribution

Urmila Mahor and Sujoy Das

Master of Computer Application,
Maulana Azad National Institute of Technology,
Bhopal, Madhya Pradesh, India

ABSTRACT: Authorship attribution tries to identify the original author of an unattributed text or document. Authorship attribution is a challenging task as it becomes quite difficult to identify original author automatically. Stylometry and authorship recognition or attribution is used interchangeably. Normally authorship attribution is done on the basis of lexical, syntactic and semantic features of a document. More recently, the problem of authorship attribution has gained wide variety attentions in the field of forensic analysis, electronic commerce etc. In this paper various feature selection, reduction and classification techniques are compared for attributing authorship of a document on PAN CLEF 2012 data set. LDA performed 12% well over all other classifiers.

KEYWORDS: machine learning, attribute selection, classification, WEKA 3.6

1 INTRODUCTION

Authorship attribution is a combination of art and science to identify the original author of an unattributed text or document. Word and sentence length features were initially used for identifying the original author of an unattributed text or document. Later many researchers followed this with some additional features like word frequencies, sentence frequency count, character frequency count, graph based approaches etc. As the work progressed in attribution, features increased in numbers and new features came in existence. In late 1990s researchers have also used classification and machine learning techniques to solve the problem. Author's intrinsic nature and his/her habitual writing style makes his/her style distinct from others and because this natural phenomena author attribution cannot be consciously manipulated. Capturing writing style by selecting features automatically is a challenging task. The features should be measurable and salient to make authors style unique. In addition, these features should be sufficient enough to distinguish authors writing style in the same topic, same genre and same periods [11]. Large number of features and irrelevant and redundant data may degrade the performance of the proposed algorithm therefore feature reduction is also an important task [4]. In this paper we have tried to find out optimal set of features to quantify and identify the writing style from the supervised set of authors, using various attribute selection and classification methods. Performance of five attribute selection methods and eight classifiers are compared.

2 LITERATURE SURVEY

Authorship attribution started in eighteen century and various researchers used different parameters for judging the authorship attribution. Initially word & sentence length were used as features. Later on researcher used word frequency count, character frequencies and function of vocabulary richness, and graph based methods as measures for studying authorship attribution.

Mingzhe et al. [19] emphasized use of comma as a feature, as it is used by an author as a break point in a sentence to clarify pause and is being used differently by different authors.

Corresponding Author: Urmila Mahor

Andrew et al. [4] used novel topic cross validation for measuring the authorship in their work. Cross validation is performed on the unseen topics of the training data set. Precision, recall F-measures were used for finding the results. Ali Osman Kausakci [5] proposed k-NNRV method as a new tool to deal with the variations in the styles. This method helps in recognizing the new informative features.

Esteban et al. [11] focused on phrase level lexical-syntactic features and graph based representation lexical features of word prefixes, word suffixes and stop words. Character features like vowel combination, vowel permutation were also used. They found that graph based representation performed better than others.

Agramon et al. [1] [2] proposed new feature systemic functional linguistics to analyze a text. In this they considered the frequencies of conjunction, modality and comment. *Systemic Functional Linguistics (*SFL) provides a base for stylistics feature selection.

Michael Gamon [20] used shallow linguistic analysis and a deep linguistic analysis features and concluded that deep linguistic analysis features in authorship attribution reduced the error rate over the function word frequencies.

2.1 FEATURES

Writing style can be identified by selecting either lexical or syntactic feature. Number of features can be identified with respect to writer but there is no particular opinion in the matter of selection of these features to be used in standard research for identifying unattributed text. Rudman [27] also mentioned that particular words may be used for a specific classification but they cannot be counted on for style analysis in general. Many kinds of tools, techniques and methodologies have been proposed for use in authorship attribution. Today's stylistic measures are based on the following features.

2.1.1 LEXICALLY-BASED METHODS

The first proposed works in authorship attribution had been focused exclusively on low-level measures such as word-length, syllables per word, and sentence-length. For these one needs tokenizer, stemmer, lemmatizer to handle such type of features. In 1887 Mendenhall proposed to use quantitative measures like average word length for authorship identification. Later on Mendenhall's work was followed by Yule and Morton [28] & they selected sentence length as feature for authorship identification.

Lexical features are based on word, sentence, or paragraph length count. Bag of word is also an approach particularly used for the selection of lexical based feature sets. According to [3] there are two main trends in lexically-based approaches: 1) Those that represent the vocabulary richness of the author and 2) those that are based on frequencies of occurrence of individual words.

The selection of the specific function words to use as features is generally based on some criteria. Various sets of function words have been proposed for English like Abbasi and Chen [5] proposed a set of 301 features, a set of 303 feature words were used by Argamon, Saric, and Stein [1] in their work, Zhao and Zobel [27] used a set of 363 function words, Koppel and Schler [21] proposed a set of 480 function words, another set of 675 words were used by Argamon, Whitelaw, Chase, Hota, Garg, and Levitan [2].

2.1.2 CHARACTER FEATURES

The character features have also been used for the authorship attribution. In this focus is on those characters which are frequently used by an author in his/her work such as quotation marks, apostrophes, comma, semicolon, upper case and lowercase characters and punctuation marks. They are counted based on per sentence or per paragraph basis. These are normally used along with lexical or syntactic feature based methods. According to Kjell [18] used character n gram feature selection with nearest neighbour and Naïve Bayes classifier.

2.1.3 SYNTACTIC FEATURES

Syntactic features are related with the construction of sentence or structure of sentence. For this we require a parser, sentence splitter or chunker to represent a sentence structure. Some researchers have shown that the use of lexical features and syntactic features together improves the performance of authorship attribution as compared to individual ones [11]. Syntactic features are noun, verb, length of noun, length of verb phrases counts, etc [1]. Koppel and Schler [14] proposed use of syntactic feature in 2003 based on syntactic errors such as sentence fragmentation, run-on sentences, mismatched tense,

etc. Karlgren and Eriksson [15] focused on model based features such as syntactic features or adverb expression and presences of clauses in the sentences for authorship attribution.

2.1.4 SEMANTIC FEATURES

McCarthy, Lewis, Dufty, and McNamara [22] described semantic measures based on synonyms and hyponyms of the words. They proposed approach to extract semantic measures from WordNet. Some researchers have also applied latentsemantic analysis and systemic functional grammar [10] for attributing the text.

FEATURE REDUCTION

Generally two methods are used to reduce the feature set

1) Feature selection in which we can use threshold methods like information gain, C2 statistic, term strength, odd ratio, weirdness coefficient, thresholding, document frequency [8][26][33][34][35].
2) Feature **extraction based on** feature clustering methods [19].

3 FEATURE SELECTION

3.1 CHI-SQUARE BASED FEATURE SELECTION

χ2 is a popular feature selection algorithm. In this term and occurrence of the class are the two events [15]. Rank is assigned to each term according to

$$X^2(\mathbb{D}, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

Where e_t =1 if the document contains term t and e_t =0 otherwise C is a random variable that takes values e_c =1 if the document is in class C and e_c =0 otherwise.

3.2 CORRELATION COEFFICIENT FEATURE SELECTION

Correlation coefficient is another approach for a pairof variables (X Y), the linear correlation coefficient r is given by

$$r = \frac{\sum_i (xi - \overline{xi})(yi - \overline{yi})}{\sqrt{\sum_i (xi - \overline{xi})^2} \sqrt{\sum_i (yi - \overline{yi})^2}}$$

Where x_i is the mean of X, and y_i is the mean of Y .The value of r lies between -1 and 1, inclusive. If X and Y are completely correlated, r takes the value of 1 or -1, if X and Y are totally independent, r is zero [7, 8].

3.3 PRINCIPAL COMPONENT ANALYSIS (PCA)

The main reason of using principle components analysis is to derive new variables that are linear combinations of the original variables. Savoy and Jacques [17] used PCA to distinguish the similarity and dissimilarity between the texts in computational terms. [8] used PCA to resolve several outstanding authorship problems.

3.4 LATENT SEMANTIC ANALYSIS

It is a well-known method for extracting the dominant features from a large data sets and for reducing the dimensionality of the data. In this corpus can be represented as a term-document matrix, which is obtained by constructing the new reductive feature space. Each document is represented by

$$d' = d^T U_k$$

Where *d'* is the new reduced feature vector and d^T is the feature vector applied by the above mentioned feature selection method.

3.5 SVM FEATURE EVALUATOR

Support Vector Machine (SVM) is well known for categorizing text. Zheng et al. [30] used SVM for authorship attribution. SVM is successful because of its good properties of regularization, maximum margin,robustness etc [31].

3.6 ONER

OneR, stands for "One Rule", it is a simple classification algorithm that generates a one-level decision tree [20].

4 LEARNING AND EVALUATION

The performance of the different methods that are studied in this research is compared by calculating precision, recall and F-measure.

Precision: Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. In term of true positive and false positive it is defined as

$$\text{Precision}=\frac{TP}{TP+FP}$$

Recall: Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. In term of true positive and false positive it is defined as

$$\text{Recall}=\frac{TP}{TP+FN}$$

F-measure is the harmonic mean of the precision and recall. F-measure focuses on the positive class, even if inverted, are devalued compared to positive features. It is defined as

$$F=\frac{2Xprecision\ X\ Recall}{(Recall+precision)}$$

5 METHODOLOGY AND EXPERIMENT

In this work various feature selection algorithms have been used to reduce the size of feature vector and then different classification algorithms are applied on the reduced feature set. Comparison is based on precision, recall and F measure. The experiment are performed on PAN CLEF 2012 data set using WEKA 3.6. MSE, RMSE, RAE, RRSE, Kappa Coefficient are used to report the errors. Total 18 features either lexical or character features are extracted from collection (Table 1.0).

Table 1.0 Feature set

Lexical Features	Character Features
a. Average length of paragraph(in line)	a. Apostrophe per para
b. Average paragraph length (in sentence)	b. Question marks per para
c. Average length of line(in words)	c. Quotation marks per paragraph
d. Average sentence length (in words)	d. Minimum character per word
e. Number of different words Complexity factor (Lexical Density)	e. Average characters per paragraph
f. Average Syllables per Word	f. Average comma per line
g. Average syllables per paragraph	
h. Average syllables per sentence	
i. Readability (Gunning-Fog Index)	
j. Readability (Alternative) beta : (100-easy 20-hard, optimal 60-70)	
k. Max sentence length in words	
l. Min sentence length in words	

The average length of paragraph (in line) is not same as that of average paragraph length (in sentence). In a line if a sentence is paused using comma then it is considered as two sentences in one line in case of average paragraph length (in sentence) otherwise line is considered as single sentence in case of average paragraph length (in sentence). The process of feature extraction for authorship attribution is shown in Figure 1. We have used five feature selection and extraction

methods and the results are shown in Table 1. All the five methods prompted us to select 8 to 10 features for attributing the authorship for PAN CLEF 2012 data set.

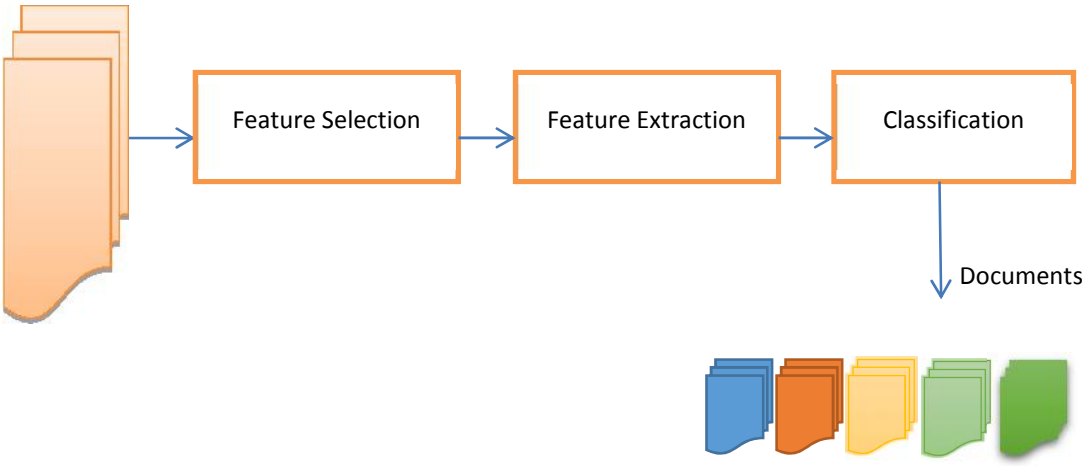

Figure 1: Process of feature extraction for authorship attribution

Table 2.0: Feature selection and extraction results

Methods	Evaluator type	No. of attribute selected out of (18)	Lexical Features as Extracted by the methods from the collection Out of (12)	Character Features as Extracted by the methods from the collection Out of(06)
Principal Components Analysis	Unsupervised	9	a,b,c,d,e,i	b,c,e
Chi-squared	Supervised	10	a,b,c,d,i,j	a,b,c,d,f
Latent Semantic Analysis	Unsupervised	8	a,b,c,g,h,i	b,c,f
SVM	Supervised	10	a,b,c,d,i,j,k	b,c,e,f
OneR	Supervised	10	b,d,e,f,h,i,j,	b,c,e

The features extraction process can be assumed to be language dependent [34]. These features are helpful in understanding the particular style of writing of an author. Most of the feature extraction techniques reported 8 to 10 features out of 18 features in this study. The extracted features are shown in Table 3.0. Eight different classifiers were applied to find out the best classification technique to classify the data. The results are shown in Table 2. LAD Tree classifier performed well as a classification technique in comparison to other techniques. The performance metrics used in these study are MSE, RMSE, RAE, RRSE ,Kappa Coefficient.

Table 3.0: Extracted Feature Set

	Lexical features	Character features
1.	Average length of paragraph(in line)	Average comma per line
2.	Average paragraph length (in sentence)	Question marks per para
3.	Average length of line(in words)	Quotation mark per para
4	Average syllables per paragraph	
5	Average syllables per sentence	
6	Readability (Gunning-Fog Index)	

Table 4.0: Comparison table based on various characteristics

Classifier	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
Hyper Pipes	51	49	0.0166	0.4962	0.4972	99.328	99.45
Conjunctive rules	53	47	0.0486	0.4867	0.5251	97.418	105.03
LAD Tree	66	34	0.3189	0.3794	0.5173	75.9483	103.47
Naïve Bayes	54	46	0.0844	0.4557	0.522	91.2149	104.39
Attribute Selector	52	48	0.0596	0.4997	0.524	100.017	104.80
CV parameter selection	52	48	0	0.4996	0.5	100	100
Meta classification via clustering	50	50	0.0048	0.5	0.7071	100.08	141.43
Lazy LWL	54	46	0.0887	0.4672	0.5167	93.52	103.33

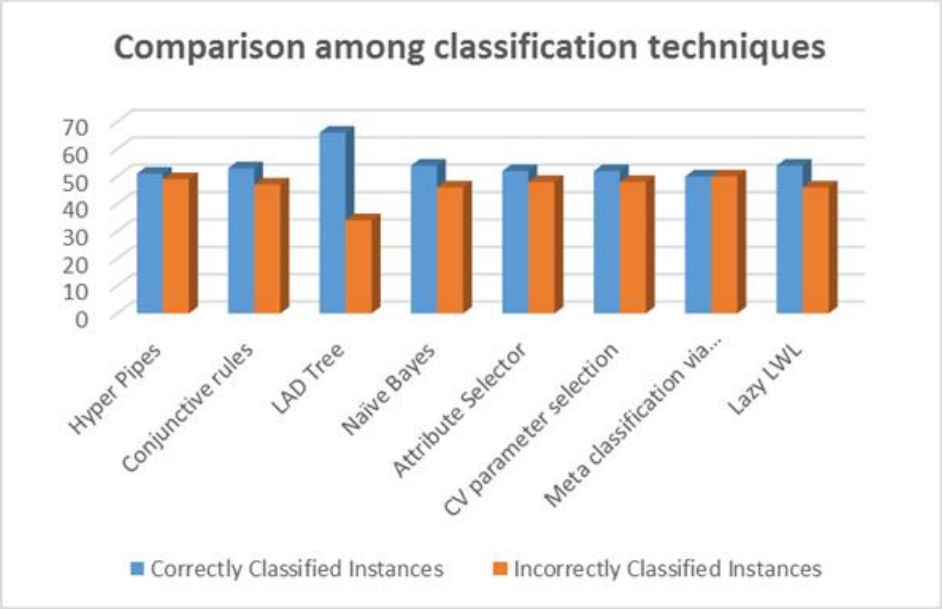

Figure 2: Comparison between correctly and incorrectly classified

Table 5.0: Comparison table based on Precision and Recall

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure
Hyper Pipes	0.51	0.526	0.46	0.51	0.38
Conjunctive rules	0.53	0.482	0.527	0.53	0.519
LAD Tree	0.66	0.341	0.66	0.66	0.66
Naïve Bayes	0.54	0.45	0.54	0.54	0.53
Attribute Selector	0.52	0.45	0.52	0.48	0.5
CV parameter selection	0.52	0.52	0.27	0.52	0.35
Meta classification via clustering	0.5	0.5	0.49	0.5	0.49
Lazy LWL	0.54	0.45	0.54	0.54	0.53

6 CONCLUSION

This study used different attribute selection algorithms in which LSA performed best as a feature selector method followed by PCA. Various classifiers are applied on PAN CLEF 2012 data to perform the experiments for estimating the classification accuracy of different classifiers. The experiments were performed on Weka3.6. PAN-CLEF authorship attribution test data set is used for performing the experiments. The eight classifiers used are Hyper-pipes, conjunctive rules, LAD Tree, Naïve Bayes, Attribute Selector, CV Parameter selection, Meta Classifier, Lazy LWL. The LAD Tree Classifier performed 12% better than rest of the techniques.

REFERENCES

[1] Argamon, S., Saric, M., & Stein, S. (2003).Style mining of electronic messages for multiple authorship discrimination. In Proceedings of the 9th ACM SIGKDD (pp. 475-480).

[2] Argamon, S., Whitelaw, C., Chase, P., Hota, S.R., Garg, N., & Levitan, S. (2007).Stylistic text classification using functional lexical features Journal of the American Society for Information Science and Technology, 58(6), 802-822.

[3] Ahmed Abbasi and Hsinchun Chen (2005). "Applying Authorship analysis to extremist group web forum messages".

[4] Andrew Y. Ng and Michael I. Jordan. 2001. On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. In Advances in Neural Information Processing Systems 14(NIPS), pages 841–848.

[5] Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems, 20(5), 67-75.d.

[6] Baayen, R., van Halteren, H., Neijt, A., & Tweedie, F. (2002).An experiment in authorship attribution.In Proceedings of JADT 2002: Sixth International Conference on Textual Data Statistical Analysis (pp. 29-37).

[7] Baayen, R., van Halteren, H., & Tweedie, F. (1996 "Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution", Literary and Linguistic Computing, 11(3), 121–131.

[8] Binongo, J. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. Chance, 16(2), 9-17

[9] Casy Whilelaw and Shlomo Agramon "Authorship Attribution using committee machines with K-nearest neighbors", rated vating Systemic Functional Features in Stylistic text classification.

[10] Efstathios Stamatatos. "A Survey of Modern Authorship Attribution Methods", Dept. of Information and Communication Systems Eng. University of the Aegean Karlovassi, Samos – 83200, Greece

[11] Esteban Castillo, Dames Vilarino, David Pinto 2012."Graph-based and lexical –syntactic approaches for authorship attribution task by Andrew I. Schin, Johnie ".

[12] F. Caver, Randale J. Honaker, and H. Martell "Authors characteristics writing styles" as seen through their use of commas Author Attribution evaluation with novel topic cross validation".

[13] Jacques Savoy .Authorship Attribution: A Comparative Study of Three Text Corpora and Three Languages .Published in Journal of Quantitative Linguistics 19, issue 2, 132-161, 2012.

[14] Koppel, M., & Schler, J. (2003). "Exploiting stylistic idiosyncrasies for authorship attribution". In Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis (pp. 69-72).

[15] Karlgren, J., & Eriksson G. (2007). Authors, genre, and linguistic convention. In Proceedings of the SIGIR Workshop on Plagiarism Analysis, Authorship Attribution, and Near-Duplicate Detection (pp. 23-28).

[16] Karel Fuka, Rudolf Hanka .Feature Set Reduction for Document Classification Problems.

[17] Jacques Savoy .Authorship Attribution: A Comparative Study of Three Text Corpora and Three Languages. Published in Journal of Quantitative Linguistics 19, issue 2, 132-161, and 2012.

[18] Kjell, B. (1994). Discrimination of authorship using visualization. Information Processing and Management, 30(1), 141-150.

[19] Jin Mingzhe ; Dept. & Grad. Sch. of Culture & Inf. Sci., Doshisha Univ., Kyoto, Japan ; Minghu Jiang Text clustering on authorship attribution based on the features of punctuations usage.

[20] Michael Gamon Linguistic correlates of style: "authorship classification with deep linguistic analysis features".

[21] Moshe Koppel, Jonathan Schler and Shlomo Argamon "Computational Methods in Authorship Attribution".

[22] McCarthy, P.M., Lewis, G.A., Dufty, D.F., & McNamara, D.S. (2006).Analyzing writing styles with coh-metrix In Proceedings of the Florida Artificial Intelligence Research Society International Conference (pp. 764-769).

[23] Shlomo Argamon,Casey Whitelaw,Paul Chase,Sushant Dhawle,Sobhan Raj Hota, Navendu Garg, Shlomo Levitan ."Stylistic Text Classification Using Functional Lexical Features".

[24] Jacques Savoy (2012): Authorship Attribution: A Comparative Study of Three Text Corpora and Three Languages, Journal of Quantitative Linguistics, 19:2,

[25] Philip M. McCarthy, Gwyneth A. Lewis, David F. Dufty, Danielle S. McNamara “Analyzing Writing Styles with Coh-Metrix”.
[26] Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. Computers and the Humanities, 31, 351-365.
[27] Ying Zhao Justin Zobel (2007).”Searching with Style: Authorship Attribution” in Classic Literature.
[28] Yule, G.U. (1938). On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. Biometrical, 30, 363-390. “Authorship attribution with thousands of candidate authors”. In Proceedings of the 29th ACM SIGIR (pp. 659-660).
[29] Zhao Y., & Zobel, J. (2005).Effective and scalable authorship attribution using function words. In Proceedings of the 2nd Asia Information Retrieval Symposium.
[30] Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing style features and classification techniques. Journal of the American Society of Information Science and Technology, 57(3), 378-393.
[31] Felipe Alonso-Atienza , José Luis Rojo-Álvarez , Alfredo Rosado-Muñoz , Juan J. Vinagre Arcadi García-Alberola , Gustavo Camps-Valls “Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection”.
[32] Ms.I.Kirubaraji, Ms.R.Jothilakshmi Comparison of Extracting Content with Minimization of Lexeme in a Text Corpus by Using Different Dimension Reduction Techniques International Journal of advanced Research in Computer and Communication Engineering Vol. 1, Issue 10, December 2012.
[33] Jacques Savoy Computer Science Department, University of Neuchatel, Rue Emile Argand 11, 2000 Neuchâtel, Switzerland “Feature Selections for Authorship Attribution”.
[34] George Forman “An Extensive Empirical Study of Feature Selection Metrics for Text Classification” Journal of Machine Learning Research 3 (2003) 1289-1305
[35] http://2012books.lardbucket.org/books/english-for-business-success/s06-01-commas.html