Proceedings of the EACL 2012 Workshop on Computational Linguistics and Writing, pages 27–34,
Avignon, France, April 23, 2012.

statistics and rules. Hence, this paper aims to contribute to the statistical approach applied to grammar checking.

The Google Books *N*-gram Corpus is a database of *n*-grams of sequences of up to 5 words and records the frequency distribution of each unit in each year from 1500 onwards. The bulk of the corpus, however, starts from 1970, and that is the year we took as a starting point for the material that we used to compile our reference corpus.

The idea of using this database as a grammar checker is to analyse an input text and detect any sequence of words that cannot be found in the *n*-gram database (which only contains *n*-grams with frequency equal to or greater than 40) and, eventually, to replace a unit in the text with one that makes a frequent *n*-gram. More specifically, we conduct four types of operations: accepting a text and spotting possible errors; inflecting a lemma into the appropriate form in a given context; filling-in the blanks in a text; and selecting, from a number of options, the most probable word form for a given context. In order to evaluate the algorithm, we applied it to solve exercises from a Spanish grammar book and also tested the detection of errors in a corpus of real errors made by second language learners.

The paper is organised as follows: we first offer a brief description of related work, and then explain our methodology for each of the experiments. In the next section, we show the evaluation of the results in comparison to the Microsoft Word grammar checker and, finally, we draw some conclusions and discuss lines of future work.

2 Related Work

Rule-based grammar checking started in the 1980s and crystallised in the implementation of different tools: papers by MacDonald (1983), Heidorn et al. (1982) or Richardson and Braden-Harder (1988) describe some of them (see Leacock et al., 2010, for a state of the art related to studies focused on language learning). This approach has continued to be used until recently (see Arppe, 2000; Johannessen et al., 2002; and many others) and is the basis of the work related with the popular grammar checker in Microsoft Word (different aspects of the tool are described in Dolan et al., 1993; Jensen et al., 1993; Gamon et al., 1997 and Heidorn, 2000: 181-207, among others). The knowledge-rich approach needs mechanisms to take into account errors within a rigid system of rules, and thus different strategies were implemented to gain flexibility (Weischedel and Black, 1980; Douglas and Dale, 1992; Schneider and McCoy, 1998 and others). Bolt (1992) and Kohut and Gorman (1995) evaluated several grammar checkers available at the time and concluded that, in general, none of the proposed strategies achieved high percentages of success.

There are reasons to believe that the limitations of rule-based methods could be overcome with statistical or knowledge-poor approaches, which started to be used for natural language processing in the late 1980s and 1990s. Atwell (1987) was among the first to use a statistical and knowledge-poor approach to detect grammatical errors in POS-tagging. Other studies, such as those by Knight and Chandler (1994) or Han et al. (2006), for instance, proved more successful than rule-based systems in the task of detecting article-related errors. There are also other studies (Yarowsky, 1994; Golding, 1995 or Golding and Roth, 1996) that report the application of decision lists and Bayesian classifiers for spell checking; however, these models cannot be applied to grammar error detection. Burstein et al. (2004) present an idea similar to the present paper, since they use *n*-grams for grammar checking. In their case, however, the model is much more complicated since it uses a machine learning approach trained on a corpus of correct English and using POS-tags bigrams as features apart from word bigrams. In addition, they use a series of statistical association measures instead of using plain frequency.

Other proposals of a similar nature are those which use the web as a corpus (Moré et al., 2004; Yin et al., 2008; Whitelaw et al., 2009), although the majority of these authors also apply different degrees of processing of the input text, such as lemmatisation, POS-tagging and chunking. Whitelaw et al. (2009), working on spell checking, are among the few who disregard explicit linguistic knowledge. Sjöbergh (2009) attempted a similar approach for grammar checking in Swedish, but with modest results. Nazar (in press) reports on an experiment where corpus statistics are used to solve a German-language multiple choice exam, the result being a score similar to that of a native speaker. The system does not use any kind of explicit knowledge of German grammar or vocabulary: answers are found by simply querying a search engine and selecting the most frequent combination of words. The present paper is a continuation and extension of that idea, now with a specific application to the practical problem of checking the grammar of texts in Spanish.

In spite of decades of work on the subject of grammar-checking algorithms, as summarised in the previous lines, the general experience with commercial grammar checkers is still disappointing, the most serious problem being that in the vast majority of cases errors in the analysed texts are left undetected. We believe that, in this context, a very simple grammar checker based on corpus statistics could prove to be helpful, at least as a complement to the standard procedures.

3 Methodology

In essence, the idea for this experiment is rather simple. In all the operations, we contrast the sequences of words as they are found in an input text with those recorded in Google's database. In the error detection phase, the algorithm will flag as an error any sequence of two words that is not found in the database, unless either of the two words is not found individually in the database, in which case the sequence is ignored. The idea is that in a correction phase the algorithm will output a ranked list of suggestions to replace each detected error in order to make the text match the *n*-grams of the database. The following subsections offer a detailed description of the methodology of each experiment. For the evaluation, we tested whether the algorithm could solve grammar exercises from a text-book (Montolío, 2000), which is one of the most widely used Spanish text-books for academic writing for native speakers, covering various topics such as pronouns, determiners, prepositions, verb tenses, and so on. In addition, for error detection we used a corpus of L2 learners (Lozano, 2009).

3.1 Error Detection

Error detection is, logically, the first phase of a grammar checking algorithm and, in practice, would be followed by some correction operation, such as those described in 3.2 to 3.4. In the error detection procedure, the algorithm accepts an input sentence or text and retrieves the frequency of all word types (of forms as they appear in the text and not the lemmata) as well as all the different bigrams as sequences of word forms, excluding punctuation signs. The output of this process is the same text with two different types of flags indicating, on the one hand, that a particular word is not found or is not frequent enough and, on the other hand, that a bigram is not frequent. The frequency threshold can be an arbitrary parameter, which would measure the "sensitivity" of the grammar checker. As already mentioned, the minimum frequency of Google *n*-grams is 40.

As the corpus is very large, there are a large number of proper nouns, even names that are unusual in Spanish. For example, in the sentence *En 1988 Jack Nicholson, Helen Hunt y Kim Basinger recibieron sendos Oscar* ('In 1988 Jack Nicholson, Helen Hunt and Kim Basinger each received one Oscar'), bigrams such as *y Kim* or, of course, others like *Jack Nicholson* are considered frequent by the system because these actors are famous in the Spanish context, but this is not the case for the bigram *Martín Fiz*, belonging to another sentence, which is considered infrequent and treated as an error (false positive), because the name of this Spanish athlete does not appear with sufficient frequency. Future versions will address this issue.

3.2 Multiple Choice Exercises

In this scenario, the algorithm is fed with a sentence or text which has a missing word and a series of possibilities from which to decide the most appropriate one for that particular context.

For instance, given an input sentence such as *El coche se precipitó por *un,una* pendiente* ('The car plunged down a slope'), the algorithm has to choose the correct option between *un* and *una* (i.e., the masculine and feminine forms of the indefinite article).

Confronted with this input data, the algorithm composes different trigrams with each possibility and one word immediately to the left and right of the target position. Thus, in this case, one of the trigrams would be *por un pendiente* and, similarly, the other would be *por una pendiente*. As in 3.1., the selection procedure is based on a frequency comparison of the trigrams in the *n*-gram database, which in this case favours the first option, which is the correct one.

In case the trigram is not found in the database, there are two back-off operations, consisting in separating each trigram into two bigrams, with the first and second position in one case and the second and third in the other. The selected option will be the one with the two bigrams that, added together, have the highest frequency.

3.3 Inflection

In this case, the exercise consists in selecting the appropriate word form of a given lemma in a given context. Thus, for instance, in another exercise from Montolío's book, *No le *satisfacer* en absoluto el acuerdo al que llegaron con sus socios alemanes* ('[He/She] is not at all satisfied with the agreement reached with [his/her] German partners'), the algorithm has to select the correct verbal inflection of the lemma *satisfacer*.

This operation is similar to the previous one, the only difference being that in this case we use a lexical database of Spanish that allows us to obtain all the inflected forms of a given lemma. In this case, then, the algorithm searches for the trigram *le * en*, where * is defined as all the inflectional paradigm of the lemma.

3.4 Fill-in the blanks

The operation of filling-in the blank spaces in a sentence is another typical grammar exercise. In this case, the algorithm accepts an input sentence such as *Los asuntos * más preocupan a la sociedad son los relacionados con la economía* ('The issues of greatest concern to society are those related to the economy'), from the same source, and suggests a list of candidates. As in the previous cases, the algorithm will search for a trigram such as *asuntos * más*, where the * wildcard in this case means any word, or more precisely, the most frequent word in that position. In the case of the previous example, which is an exercise about relative pronouns, the most frequent word in the corpus and the correct option is *que*.

4 Results and Evaluation

4.1 Result of error detection

The results of our experiments are summarised in Table 1, where we distinguish between different types of grammar errors and correction operations. The table also offers a comparison of the performance of the algorithm against Microsoft Word 2007 with the same dataset. In the first column of the table we divide the errors into different types as classified in Montolío's book. Performance figures are represented as usual in information retrieval (for details, see Manning et al., 2008): the columns represent the numbers of true positives (tp), which are those errors that were effectively detected by each system; false negatives (fn) referring to errors that were not detected, and false positives (fp), consisting in those cases that were correct, but which the system wrongly flagged as errors. These values allowed us to define precision (P) as $tp/(tp+fp)$, recall (R) as $tp/(tp+fn)$ and $F1$ as $2.P.R/(P+R)$.

The algorithm detects (with a success rate of 80.59%), for example, verbs with an incorrect morphology, such as **apreto* (instead of *aprieto*, 'I press'). Nevertheless, the system also makes more interesting detections, such as the incorrect selection of the verb tense, which requires information provided by the context: *Si os vuelve a molestar, no *volved a hablar con él* ('If [he] bothers you again, do not talk to him again'). In this sentence, the correct tense for the second verb is *volváis*, as the imperative in negative sentences is made with the subjunctive. In the same way, it is possible to detect incorrect uses of the adjective *sendos* ('for each other'), which cannot be put after the noun, among other particular constraints: combinations such as **los sendos actores* ('both actors') or **han cerrado filiales sendas* ('they have closed both subsidiaries') are marked as incorrect by the system.

In order to try to balance the bias inherent to a grammar text-book, we decided to replicate the experiment with real errors. The decision to extract exercises from a grammar book was based on the idea that this book would contain a diverse sample of the most typical mistakes, and in this sense it is representative. But as the examples given by the authors are invented, they are often uncommon and unnatural, and of course this frequently has a negative effect on performance. We thus repeated the experiment using sentences from the CEDEL2 corpus (Lozano, 2009), which is a corpus of essays in Spanish written by non-native speakers with different levels of proficiency.

For this experiment, we only used essays written by students classified as "very advanced". We extracted 65 sentences, each containing one error.

	This Experiment						Word 2007					
Type of error	**tp**	**fn**	**fp**	**% P**	**% R**	**% F1**	**tp**	**fn**	**fp**	**% P**	**% R**	**% F1**
gerund	9	8	9	50	52.94	51.42	9	8	1	**90**	52.94	**66.66**
verb morphology	54	17	13	80.59	76.05	78.25	60	11	3	**95.23**	**84.50**	**89.54**
numerals	4	9	7	36.36	30.76	33.32	6	7	0	**100**	**46.15**	**63.15**
grammatical number	10	8	1	90.90	55.55	68.95	10	8	1	90.90	55.55	68.95
prepositions	25	40	17	59.52	**38.46**	**46.72**	13	52	0	**100**	20	33.33
adjective “sendos”	5	0	1	83.33	**100**	**90.90**	1	4	0	**100**	20	33.33
various	55	52	52	51.40	**51.40**	**51.40**	33	74	10	**76.74**	30.84	43.99
total	162	134	100	61.83	**54.72**	58.05	132	164	15	**89.79**	44.59	**59.58**

Table 1: Summary of the results obtained by our algorithm in comparison to Word 2007

Since the idea was to check grammar, we only selected material that was orthographically correct, any minor typos being corrected beforehand. In comparison with the mistakes dealt with in the grammar book, the kind of grammatical problems that students make are of course very different. The most frequent type of errors in this sample were gender agreement (typical in students with English as L1), lexical errors, prepositions and others such as problems with pronouns or with transitive verbs, among others.

Results of this second experiment are summarised in Table 2. Again, we compare performance against Word 2007 on the same dataset. In the case of this experiment, lexical errors and gender agreement show the best performance because these phenomena appear at the bigram level, as in **Después del boda* (‘after the wedding’) which should be feminine (*de la boda*), or **una tranvía eléctrica* (‘electric tram’) which should be masculine (*un tranvía*). But there are other cases where the error involves elements that are separated from each other by long distances and of course will not be solved with the type of strategy we are discussing, as in the case of **un país donde el estilo de vida es avanzada* (‘a country with an advanced lifestyle’), where the adjective *avanzada* is wrongly put in feminine when it should be masculine (*avanzado*), because it modifies a masculine noun *estilo*.

In general, results of the detection phase are far from perfect but at least comparable to those achieved by Word in these categories. The main difference between the performance of the two algorithms is that ours tends to flag a much larger number of errors, incurring in many false positives and severely degrading performance. The behaviour of Word is the opposite, it tends to flag fewer errors, thus leaving many errors undetected. It can be argued that, in a task like this, it is preferable to have false positives rather than false negatives, because the difficult part of producing a text is to find the errors. However, a system that produces many false positives will lose the confidence of the user. In any case, more important than a difference in precision is the fact that both systems tend to detect very different types of errors, which reinforces the idea that statistical algorithms could be a useful complement to a rule-based system.

4.2 Result of multiple choice exercise

The results of the multiple choice exercise in the book are shown in Table 3. Again, we compared performance with that achieved by Word. In order to make this program solve a multiple choice exercise we submitted the different possibilities for each sentence and checked whether it was able to detect errors in the wrong sentences and leave the correct ones unflagged.

Results in this case are similar in general to those reported in Section 4.1. An example of a correct trial is with the fragment **el,la* génesis del problema* (‘the genesis of the problem’), where the option selected by the algorithm is *la génesis* (feminine gender). In contrast, it is not capable of giving the correct answer when the context is very general, such as in **los,las* pendientes son uno de los complementos más vendidos como regalo* (‘Earrings are one of the accessories most frequently sold as a gift’), in which the words to choose from are at the beginning of the sentence and they are followed by *son* (‘they are’), which comes from *ser*, perhaps the most frequent and polysemous Spanish verb. The correct answer is *los* (masculine article), but the system offers the incorrect *las* (feminine) because of the polysemy of the word, since *las pendientes* also exist, but means ‘the slopes’ or even ‘the ones pending’.

	This Experiment						Word 2007					
Type of error	**tp**	**fn**	**fp**	**% P**	**% R**	**% F1**	**tp**	**fn**	**fp**	**% P**	**% R**	**% F1**
gender agreement	9	6	3	75	**60**	**66.66**	7	8	0	**100**	46.66	63.63
lexical selection	16	10	4	80	**61.53**	**69.56**	4	22	0	**100**	15.38	26.66
prepositions	2	11	2	**50**	**15.38**	**23.52**	0	13	0	0	0	0
various	4	7	5	44.44	**36.36**	**39.99**	3	8	3	**50**	27.27	35.29
total	31	34	17	64.58	**47.69**	**54.86**	14	51	3	**82.35**	21.53	34.14

Table 2: Replication of the experiment with a corpus of non-native speakers (CEDEL2, Lozano, 2009)

	Trials	This Experiment		Word 2007	
Type of error		**Correct**	**% P**	**Correct**	**% P**
adverbs	9	8	**88.89**	5	55.55
genre	10	7	**70.00**	3	30
confusion DO-IO	4	2	50.00	2	50

Table 3: Solution of the multiple choice exercise

4.3 Result of inflection exercise

Results in the case of the inflection exercise are summarised in Table 4. When giving verb forms, results are correct in 66.67% of the cases. For instance, in the case of *La mayoría de la gente *creer* que...* ('The majority of people think that...'), the correct answer is *cree*, among other possibilities such as *creen* (plural) or *creía* (past). But results are generally unsuccessful (22.22%) when choosing the correct tense, such as in the case of *Si el problema me *atañer* a mí, ya hubiera hecho algo para remediarlo* ('If the problem was of my concern, I would have already done something to solve it'). In this example, the correct verb tense is *atañera* or *atañese*, both of which are forms for the third person past subjunctive used in conditional clauses, but the system gives *atañe*, a correct form for the verb *atañer* that, nevertheless, cannot be used in this sentence. As it can be seen, the problem is extremely difficult for a statistical procedure (there are around 60 verb forms in Spanish), and this may explain why the results of this type of exercise were more disappointing.

Type of error	**Trials**	**Correct**	**% P**
verb number	9	6	66.67
verb tense	9	2	22.22

Table 4: Results of the inflection exercise

4.4 Result of filling-in the blanks

When asked to restore a missing word in a sentence, the algorithm is capable of offering the correct answer in cases such as *El abogado * defendió al peligroso asesino...* ('The lawyer -who- defended the dangerous murderer...'), where the missing word is *que*. Other cases were not solved correctly, as the fragment * *ácida manzana* ('the acid apple'), because the bigram *la ácida* is much less frequent than *lluvia ácida*, 'acid rain', the wrong candidate proposed by the system. Results of this exercise are summarised in Table 5.

Type of error	**Trials**	**Correct**	**% P**
articles	7	4	57.14
pronouns	7	3	42.86

Table 5: Results of the fill-in-the-blank exercise

5 Conclusions and Future Work

In the previous sections we have outlined a first experiment in the detection of different types of grammar errors. In summary, the algorithm is able to detect difficult mistakes such as **informes conteniendo* (instead of *informes que contenían* 'reports that contained': a wrong use of the gerund) or **máscaras antigases* (instead of *máscaras antigás* 'gas masks', an irregular plural), which are errors that were not detected by MS Word.

One of the difficulties we found is that, despite the fact that the corpus used is probably the most extensive corpus ever compiled, there are bigrams that are not present in it. This is not surprising, since one of the functions of linguistic competence is the capacity to represent and make comprehensible strings of words which have never been produced before. Another problem is that frequency is not always useful for detecting mistakes, because the norm can be very separated from real use. An example of this is that, in one of the error detection exercises, the system considers that the participle *freídos* ('fried') is incorrect because it is not in the corpus, but the participle is actually correct, even when the majority of speakers think that only the irregular form (*frito*) is normative. The opposite is also true: some incorrect structures are very frequently used and many speakers perceive them as correct, such as *ayer noche* instead of *ayer por la noche* ('last night'), or some very common Gallicisms such as **medidas a tomar* instead of *medidas por tomar* 'measures to be taken', or **asunto a discutir* ('matter to discuss') which should be *asunto para discutir*.

Several ideas have been put forward to address these difficulties in future improvements to this research, such as the use of trigrams and longer *n*-grams instead of only bigrams for error detection. POS-tagging and proper noun detection are also essential. Another possibility is to complement the corpus with different Spanish corpora, including press articles and other sources. We are also planning to repeat the experiment with a new version of the *n*-gram database this time not as plain word forms but as classes of objects such that the corpus will have greater power of generalisation. Following another line of research that we have already started (Nazar and Renau, in preparation), we will produce clusters of words according to their distributional similarity, which will result in a sort of Spanish taxonomy. This can be accomplished because all the words that represent, say, the category of vehicles are, in general, very similar as regards their distribution. Once we have organised the lexicon of the corpus into categories, we will replace those words by the name of the category they belong to, for instance, PERSON, NUMBER, VEHICLE, COUNTRY, ORGANISATION, BEVERAGE, ANIMAL, PLANT and so on. By doing this, the Google *n*-gram corpus will be useful to represent a much more diverse variety of *n*-grams than those it actually contains. The implications of this idea go far beyond the particular field of grammar checking and include the study of collocations and of predicate-argument structures in general. We could ask, for instance, which are the most typical agents of the Spanish verb *disparar* (to shoot). Searching for the trigram *los * dispararon* in the database, we can learn, for instance, that those agents can be *soldados* (soldiers), *españoles* (Spaniards), *guardias* (guards), *policías* (policemen), *cañones* (cannons), *militares* (the military), *ingleses* (the British), *indios* (indians) and so on. Such a line of study could produce interesting results and greatly improve the rate of success of our grammar checker.

Acknowledgments

This research has been made possible thanks to funding from the Spanish Ministry of Science and Innovation, project: "Agrupación semántica y relaciones lexicológicas en el diccionario", lead researcher J. DeCesaris (HUM2009-07588/FILO); APLE: "Procesos de actualización del léxico del español a partir de la prensa", 2010-2012, lead researcher: M. T. Cabré (FFI2009-12188-C05-01/FILO) and Fundación Comillas in relation with the project "Diccionario de aprendizaje del español como lengua extranjera". The authors would like to thank the anonymous reviewers for their helpful comments, Cristóbal Lozano for providing the non-native speaker corpus CEDEL2, Mark Andrews for proofreading, the team of the CIBER HPC Platform of Universitat Pompeu Fabra (Silvina Re and Milton Hoz) and the people that compiled and decided to share the Google Books *N*-gram corpus with the rest of the community (Michel et al., 2010).

References

Antti Arppe. 2000. Developing a Grammar Checker for Swedish. *Proceedings of the Twelfth Nordic Conference in Computational Linguistics. Trondheim, Norway*, pp. 5–77.

Eric Steven Atwell. 1987. How to Detect Grammatical Errors in a Text without Parsing it. *Proceedings of the Third Conference of the European Association for Computational Linguistics, Copenhagen, Denmark*, pp. 38–45.

Philip Bolt. 1992. An Evaluation of Grammar-Checking Programs as Self-help Learning Aids for Learners of English as a Foreign Language. *Computer Assisted Language Learning*, 5(1):49–91.

Jill Burstein, Martin Chodorow, Claudia Leacock. 2004. Automated Essay Evaluation: the Criterion Writing Service. *AI Magazine*, 25(3):27–36.

William B. Dolan, Lucy Vanderwende, Stephen D. Richardson. 1993. Automatically Deriving Structured Knowledge Base from On-Line Dictionaries. *Proceedings of the Pacific ACL. Vancouver, BC.*

Shona Douglas, Robert Dale. 1992. Towards robust PATR. *Proceedings of the 15th International Conference on Computational Linguistics, Nantes*, pp. 468–474.

Michael Gamon, Carmen Lozano, Jessie Pinkham, Tom Reutter. 1997. Practical Experience with Grammar Sharing in Multilingual NLP. *Proceedings of the Workshop on Making NLP Work. ACL Conference, Madrid*.

Andrew Golding. 1995. A Bayesian Hybrid Method for Context Sensitive Spelling Correction. *Proceedings of the Third Workshop on Very Large Corpora*, pp. 39–53.

Andrew Golding, Dan Roth. 1996. Applying Winnow to Context Sensitive Spelling Correction. *Proceedings of the International Conference on Machine Learning*, pp. 182–190.

Na-Rae Han, Martin Chodorow, Claudia Leacock. 2006. Detecting Errors in English Article Usage by non-Native Speakers. *Natural Language Engineering*, 12(2), pp. 115–129.

George E. Heidorn. 2000. Intelligent writing assistance. *In Dale, R, Moisl H, Somers H, eds. Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text. New York: Marcel Dekker*.

George E. Heidorn, Karen Jensen, Lance A. Miller, Roy J. Byrd, Martin Chodorow. 1982. The EPISTLE text-critiquing system. *IBM Systems Journal*, 21, pp. 305–326.

Karen Jensen, George E. Heidorn, Stephen Richardson, eds. 1993. *Natural Language Processing: The PNLP Approach*. Kluwer Academic Publishers.

Jane Bondi Johannessen, Kristin Hagen, Pia Lane. 2002. The Performance of a Grammar Checker with Deviant Language Input. *Proceedings of the 19th International Conference on Computational Linguistics. Taipei, Taiwan*, pp. 1–8.

Kevin Knight, Ishwar Chandler. 1994. Automated Postediting of Documents. *Proceedings of National Conference on Artificial Intelligence, Seattle, USA*, pp. 779–784.

Gary F. Kohut, Kevin J. Gorman. 1995. The Effectiveness of Leading Grammar/Style Software Packages in Analyzing Business Students' Writing. *Journal of Business and Technical Communication*, 9:341–361.

Claudia Leacock, Martin Chodorow, Michael Gamon, Joel Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. USA: Morgal and Claypool.

Cristóbal Lozano. 2009. CEDEL2: Corpus Escrito del Español L2. *In: Bretones Callejas, Carmen M. et al. (eds) Applied Linguistics Now: Understanding Language and Mind. Almería: Universidad de Almería. Almería*, pp. 197–212.

Nina H. Macdonald. 1983. The UNIX Writer's Workbench Software: Rationale and Design. *Bell System Technical Journal*, 62, pp. 1891–1908.

Chistopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, Erez Lieberman Aiden. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331(6014), pp. 176–182.

Estrella Montolío, ed. 2000. *Manual práctico de escritura académica*. Barcelona: Ariel.

Joaquim Moré, Salvador Climent, Antoni Oliver. 2004. A Grammar and Style Checker Based on Internet Searches. *Proceedings of LREC 2004, Lisbon, Portugal*.

Rogelio Nazar. In press. Algorithm qualifies for C1 courses in German exam without previous knowledge of the language: an example of how corpus linguistics can be a new paradigm in Artificial Intelligence. *Proceedings of Corpus Linguistics Conference, Birmingham, 20-22 July 2011*.

Rogelio Nazar, Irene Renau. In preparation. A cocurrence taxonomy from a general language corpus. *Proceedings of the 15th EURALEX International Congress, Oslo, 7-11 August 2012*.

Stephen Richardson, Lisa Braden-Harder. 1988. The Experience of Developing a Large-Scale Natural Language Text Processing System: CRITIQUE. *Proceedings of the Second Conference on Applied Natural Language Processing (ANLC '88). ACL, Stroudsburg, PA, USA*, pp. 195–202.

David Schneider, Kathleen McCoy. 1998. Recognizing Syntactic Errors in the Writing of Second Language Learners. *Proceedings of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics, Montreal, Canada*, pp. 1198–1204.

Jonas Sjöbergh. 2009. *The Internet as a Normative Corpus: Grammar Checking with a Search Engine*. Techical Report, Dept. of Theoretical Computer Science, Kungliga Tekniska Högskolan.

Ralph M. Weischedel, John Black. 1980. Responding-to potentially unparseable sentences. *American Journal of Computational Linguistics*, 6:97–109.

Casey Whitelaw, Ben Hutchinson, Grace Y. Chung, Gerard Ellis. 2009. Using the Web for Language Independent Spell Checking and Autocorrection. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore*, pp. 890–899.

David Yarowsky. 1994 Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. *Proceedings of the ACL Conference*, pp. 88–95.

Xing Yin, Jiangfeng Gao, William B. Dolan. 2008. A Web-based English Proofing System for English as a Second Language Users. *Proceedings of the 3rd International Joint Conference on Natural Language Processing, Hyderabad, India*, pp. 619–624.

Google Ngram search is a power search for Google Books. A Ngram is a statistical analysis of text to find some sort of item in the text.Â Choose a corpus. You can search foreign language texts or English texts, and in addition to the standard choices, you may notice entries such as "English (2009)" or "American English (2009)" at the bottom of the list. These are older corpora that Google has since updated, but you may have some reason to make your comparisons against old data sets. Most users can ignore them and focus on the most recent corpora. Set the smoothing level. Smoothing refers to how smooth the graph is at the end. from Wikipedia: The Google Ngram Viewer is a phrase-usage graphing tool which charts the yearly count of selected n-grams (letter combinations)[n] or words and phrases, as found in over 5.2 million books digitized by Google Inc (up to 2008). The words or phrases (or ngrams) are matched by...Â This item contains the Google ngram data for the English (fiction) languageset. Here are the datasets backing the Google Books Ngram Viewer. These datasets were generated in July 2009; we will update these datasets as our book scanning continues, and the updated versions will have distinct and persistent version identifiers (20090715 for the current set). Each of the numbered links below will directly download a fragment of the given corpus. For instance, the first ten links Books Ngram Viewer. Share Download raw data. Share.Â Inflections shook_INF drive_VERB_INF. Arithmetic compositions (color /(color + colour)). Corpus selection I want:eng_2019. Close View All options. 1800 - 2019 arrow_drop_down.