Automating the Measurement of Linguistic Features to Help Classify Texts as Technical

Terry Copeck[1], Ken Barker[2¥], Sylvain Delisle[3] & Stan Szpakowicz[1]

[1] School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, Canada, K1N 6N5
{terry, szpak}@site.uottawa.ca

[3] Département de mathématiques et d'informatique
Université du Québec à Trois-Rivières
Trois-Rivières, Québec, Canada, G9A 5H7
Sylvain_Delisle@uqtr.uquebec.ca

[2] Department of Computer Sciences
University of Texas at Austin
Austin, Texas, USA, 78712
kbarker@cs.utexas.edu

Abstract

Text classification plays a central role in software systems which perform automatic information classification and retrieval. Occurrences of linguistic feature values must be counted by any mechanism that classifies or characterizes natural language text by topic, style, genre or, in our case, by the degree to which a text is *technical*. We discuss the methodology and key details of the feature value extraction process, paying attention to fast and reliable implementation. Our results are mixed but support continued investigation—while a significant level of automation has been achieved, the successfully extracted feature counts do not always correlate with technicality as strongly as anticipated.

---

1. Introduction

Software systems that perform automatic information classification and retrieval rely crucially on text classification. Internet applications have caused this well-known, difficult problem to become an active research topic in machine learning (ML) and in computational linguistics (CL). Relevant CL work ranges from text filtering for data extraction and information retrieval (Lewis & Tong, 1992) to information classification or document filtering based on features extracted via shallow parsing (for example, Chandrasekar & Srinivas, 1997; Ikeda *et al*., 1998). The ML community has used WordNet (Miller, 1990) to support rule induction systems for text classification (for example, Junker & Abecker, 1997). ML researchers also attempt text classification or categorization based on such techniques as k-nearest-neighbour, decision trees, neural networks, or hierarchical Bayesian clustering.

The work presented here is part of an ongoing study into the linguistic nature of *technical text*, so we favour CL rather than ML techniques. Although it is widely used, the term "technical text" lacks a generally accepted definition. In one of the informal uses, for example, literal writing may be referred to as "technical" in contrast with figurative writing, with which CL systems deal very seldom. For some years we have worked on characterizing

---

¥ Work done while the author was at the School of Information Technology and Engineering, University of Ottawa.

technical texts in terms of linguistic features that can be measured semi-automatically. Such a characterization differs substantially from what is normally understood as text classification. It could, however, offer a reliable data filter to systems that claim to work for unconstrained texts but in fact are meant for texts which can be said to be technical. Success in this work would also suggest that more reliable, general text classifiers are feasible if one considers surface linguistic features more complex than just word occurrence or co-occurrence.

2. Related Work

Within the broad realm of computational linguistics, the machine translation community has recently turned to text categorization and the feature extraction it usually entails (Trosborg, 1999; Niske, 1998). Sinclair and Ball (1996) tabulate the classification systems used in 30 corpora from various languages according to literary genre, topic, medium, fiction/non-fiction, style or other feature, for example, publication type.

Some researchers apply ML techniques to word vectors. Cohen (1999) contrasts two ML algorithms that employ word context; Joachims (1999) uses transducing Support Vector Machines on word stem vectors. Word features figure in a number of systems (Papka & Allen, 1998; Larkey & Croft, 1996). Liddy *et al.* (1994) categorize texts exclusively by semantic codes assigned to words from a machine-readable dictionary. Apté *et al.* (1994) use topic-specific dictionaries, Scott & Matwin (1998) use WordNet hypernyms. Statistical approaches are also popular (Wilbur, 1996; see Yang, 1999 for an overview). Common elements in these systems are the primacy of the word in some form, a presumption of complete automation, and (in most cases) the absence of a pre-existing taxonomy. The latter means that documents themselves provide the characteristics that classify them best.

Froissart and Lallich-Boidin (1999) discuss how to determine features of technical texts in the context of information retrieval. They identify good indicators of technicality (for example, well identified readership) and some others that discriminate less well (for example, types of anaphora). Many of these features, rooted in semantic and even pragmatic aspects of text, would be quite difficult to automate. The authors conclude that a better characterization of text technicality is useful for information retrieval purposes, as well as for other natural language analysis tasks, by allowing the most appropriate processing strategies for such texts.

Our approach is most akin to that of Douglas Biber, who has worked on text typology for many years. He identifies 67 linguistic features (1988) upon which to classify text, mostly unambiguous surface phenomena such as nominalizations, passives, contractions etc. Biber worked with the LOB and London-Lund corpora; Karlgren and Cutting (1994) applied Biber's technique to the Brown corpus, concentrating on automatability and selecting only features easy to compute mechanically for a given text. Biber employed factor analysis to classify texts into the most distinct categories possible; Karlgren and Cutting constructed functions to discriminate among the texts in the existing Brown corpus categories, grouped into two or four more general classes or taken as-is. Discriminant functions were the more successful classifiers when fewer categories were involved. Karlgren and Cutting suggest content-based classification together with filtering based on lower-level cues.

While our work aims at a predictive rather than descriptive characterization of texts, Karlgren and Cutting's experience parallels ours. The differences noted in the preceding paragraph notwithstanding, our project is consistent with Biber's approach and can be seen as its direct successor. Kilgarriff's review (1995) of Biber's research program notes that "as yet the methodology has only been used by Biber and a small group of collaborators" because it is "technically difficult and time-consuming to implement", but "this is the only such machinery there is". Illouz *et al.* (2000) also present work based on Biber's research—a prototype "text profiling" system. It aims to compute a measure of homogeneity of a corpus, relative to some parameters (for example, to certain grammatical features).

3. Selecting Text Features

Our work on characterizing texts as technical began by identifying lexical, syntactic, semantic and pragmatic features thought likely to indicate technicality. Copeck *et al.* (1997) describe an experiment in which 20 raters consider nine texts and assess them in terms of 42 such features and overall technicality on a five-point scale. Included in the nine texts were what had been expected to be positive and negative examples, for example, an academic paper and a poem. Spearman correlations were computed between feature values and the overall technicality rating across texts. A formula for computing a text's technicality was constructed by solving a set of linear equations whose variables were those features correlating significantly with overall technicality. This is provisional; the question of how to use these results in a methodology to classify texts remains under investigation.

To accommodate the 5-point ordinal scale of technicality, ordinal regression analysis was attempted. A latent variable approach allows us to model technicality categories as a function of the data. The logistic and normal latent variable regression methods were used, with model fitting based on the PROC LOGISTIC procedure in SAS 6.11 (SAS Institute, 1996). Below, Table 1 reports χ^2 results for the 15 most significant features with the probability that a feature's parameter is greater than χ^2 appearing in brackets. Both methods rank features in much the same order. A stepwise regression analysis based on the same two methods recovered 10 of these 15 variables, including the top 5 most significant variables. As a further validation, we fit a Bayesian ordinal regression model similar to Albert and Chib (1993) using Markov chain Monte Carlo. This gave results very similar to those in Table 1.

To this point, candidate features had been chosen for possible correlation with technicality, not for ease of measurement. Attention then turned to applications, which require automatic measurement. 17 of the 42 features were identified as wholly or partially automatable, and the normal and proportional-odds latent variable regression analysis was redone on data for these features alone. Ten proved to correlate sufficiently with technicality to warrant further investigation. Adding the constraint of measurement automation, however, is likely to weaken any technicality formula because it will eliminate features indicative of technicality which can only be assessed by humans.

While automating feature measurement excludes some strongly correlated but subjective features (such as *serious treatment of the subject matter*), it allows addition of features not

Feature	Logistic R.A.	Normal R.A.
Topic Identified	18.2 (0.0001)	20.4 (0.0001)
Verbs In Present Tense	14.9 (0.0001)	14.5 (0.0001)
Serious Treatment	8.7 (0.0031)	9.4 (0.0021)
Use Of Conventions	7.2 (0.0071)	8.4 (0.0038)
Colloquialisms	6.2 (0.0127)	6.0 (0.0141)
Increasing Complexity	5.6 (0.0173)	6.5 (0.0106)
Meaning By Denotation	4.7 (0.0286)	5.7 (0.0163)
Communicates Knowledge	4.5 (0.0323)	5.2 (0.0220)
Unambiguous References	4.0 (0.0455)	5.8 (0.0159)
Introduction Present	3.2 (0.0717)	2.9 (0.0885)
Ellipsis	2.9 (0.0837)	3.2 (0.0725)
Use Of Terminology	2.7 (0.0970)	3.1 (0.0801)
Single-Sense Domain Verbs	2.7 (0.0987)	2.6 (0.1068)
Discourse Particles	1.5 (0.2139)	1.6 (0.2111)
Passive Voice	1.3 (0.2425)	1.4 (0.2356)

Table 1: χ^2 values and ρ-values (in brackets) for 15 most significant variables from proportional-odds and normal latent variable ordinal regression.

possible in experiments with human judges (such as *number of unique words*). We added six new features; they appear at the bottom of Table 2 below.

A new experiment had two phases. First, we developed a set of automatic feature count techniques. Next, we applied these techniques to the selected features in texts of known technicality to see whether it is indicated correctly. Texts were chosen from the Brown corpus (Francis & Kucera, 1982) which includes a variety of texts of greater and lesser technicality. We chose the Brown corpus because it does not need any special introductions in the CL community, and because it represents a wide variety of contemporary language usage, technical and non-technical in the intuitive sense. Eleven judges rated the technicality of thirty texts chosen equally from Brown's fifteen categories; average ratings ranged from 1.13 to 4.75 on a scale of 1 (low) to 5 (high).

The agreement among raters, expressed as the multi-rater kappa statistic (Siegel & Castellan, 1988; Carletta, 1996), was rather low. We calculated it with the SPSS *mkappasc* procedure, on 11 raters' technicality ratings of 30 Brown texts chosen nearly equally from Brown's 15 genres. Some sources set the kappa threshold for *high agreement* at the rather steep 0.7 to 0.8, but lower thresholds are also considered. Our result was 0.33, which may be tolerable when applying *mkappasc* to five-point scale data rather than, as is usual, to a set of binary decisions.

4. Semi-Automatic Acquisition of Feature Values

We now discuss the automation of measuring these features in the 500 texts in the Brown corpus. The strategy was first to measure each feature as automatically as possible, postponing any manual processing. For 14 of the 16 features such automatic measurement was feasible. Manual intervention was required to detect *conventions* and *colloquialisms*. Templates to recognize the telltale format of such items as stock market prices, sports scores, or recipe ingredient lists can be constructed. We are, however, unaware of any catalogue of such templates, and coding each would be as much work as identifying any single textual feature. *Conventions* were therefore counted by hand. Because the use of words in a non-literal sense is only evident to a human reader, *colloquialisms* were also assessed entirely manually.

No attempt was made to determine the precision and recall of feature measurement. The effort required to determine exact counts of the features involved was beyond the scope of the

Name	Description
Colloquialisms	use of colloquial words and phrases
Present Tense	verbs in present tense
Conventions	use of a mutually-understood convention
Citations	citation of authorities
Title, Headings	title or headings in the text
Introduction	presence of an introductory section
Passives	verbs in passive voice
Examples	explicitly identified examples
Interrogatives	questions
Binders	use of discourse connectives
Word Length	average length of a word in characters
Sentence Length	average length of a sentence in words
Paragraph Length	average length of a paragraph in words
Unique Words	number of unique words
Vocabulary Words	number of dictionary words
Vocabulary Familiarity	frequency list rank of vocabulary words

Table 2: Automatable feature indicators of text technicality

experiment, as it had been for Biber, and for Karlgren and Cutting. Moreover, such measures are less important to us than they might have been to those authors. The accuracy of a mechanism that predicts a text's technicality, which we hope ultimately to derive in this work, is what should validate our approach. In any event, automatic processing was biased towards recall rather than precision, in order not to miss instances. That is to say, filters were designed to allow false positives to pass in order not to miss true negatives. The former were then eliminated in the manual inspection that followed.

To measure the 14 fully automatable features, we used several public domain text processing resources: Brill's tagger (Brill 1992), the Collins part-of-speech (POS) dictionary, Marcu's list of 461 discourse connectives (Marcu 1998), an additional SGML-encoded version of Brown, and a word frequency list derived from four large general-purpose text corpora (Copeck *et al.* 1999). Customized versions of some of these programs were created: a silent version of Brill's parser accepts a file of text from the command line. A shell script packages access to the consolidated frequency list, which was restructured into a format suitable for grep: *word* and *rank* separated by a space. In general, feature processing strategies involved two or three steps; typically, a perl program or grep pattern would be applied to each file in the corpus by a shell script and output accumulated in a log file for subsequent processing. This occurred often enough to warrant the writing of the 'runner' script. While individually trivial, such small programs and scripts performed so much of the work in this experiment as to require mention. Log files were then imported into a spreadsheet where, after an optional manual editing pass, data was summed for each text and averaged for all texts in a genre to produce the results reported in Section 5.

Resources for the work were then assembled. Problems in our copy of Brown were corrected (71 circumflexes marking accents were deleted, and 687 errors consisting mostly of run-together words were fixed). Six supplementary files, with the name suffixed *.sgml*, *.tag*, *.unk*, *.freq*, *.voc* and *.faml*, were then constructed for each Brown text. Tags were stripped off the SGML version of Brown, which has sentences marked, producing a one-sentence-per-line, two-return-per-paragraph version (*.sgml*). The six low-level features could easily be extracted from this version. The corpus was tagged with Brill's tagger, providing a POS-marked version of each text (*.tag*). A list of unknown words, not matched in the Collins dictionary, was produced for each text (*.unk*), a frequency list constructed for all its words (*.freq*) and a vocabulary list (*.voc*) assembled from those matched in Collins. Finally, a measure of familiarity was computed for each word in the vocabulary frequency list (*.faml*).

The familiarity measure, named to suggest WordNet's *faml* search, associates a vocabulary word with its rank in the consolidated frequency list. Because of ties, this list of 60,068 words distinguishes 47,976 ranks; for example, the word *to* has rank 1, while *flyers* and *twinkling* share rank 14,144. The closer a word's rank is to the top, the more commonly it appears in the four large general-purpose corpora from which the frequency list is derived.

We now present details of the automated and manual processing of each feature. Though developing counts is a simple concept and in practice has proven easy for some features, for others it is not trivial, and in certain cases we were unable to automate the core process. In the discussion, a *line* of text contains a sentence or a title, which usually is a phrase. Tags are from Brill's system. Whenever meaningful, the percentage of automatically-extracted instances that passed the subsequent manual inspection are given.

Colloquialisms. Instances of colloquial phrases were extracted from each text file manually and its file of non-vocabulary words (.unk) was scanned in parallel for single colloquial words. Processing was primarily manual because there is no mechanism that can automatically identify colloquial usages in text. Included in the count were onomatopoeia like *rat-a-tat-tat* or coinages like *Americanegro*, but not figures of speech in general.

Present Tense. The .tag file, produced by Brill's tagger, was searched for tag sequences that may be given to present tense constructions. Three types of sequences were considered: /MD + "be" + "-ing"/VBG, (/VBZ || /VBP) + "-ing"/VBG, /VB + not (/VB || /VBD || /VBG || /VBN). These three patterns match such verb phrases as, respectively, *must be writing*, *am writing*, *write* + not a verb. As a check, tokens in the extraction were looked up in the Collins dictionary and those with no 'V' POS entry were deleted. Global searches were performed for *might*, *could*, *should*, *would*, *will*, *shall*, *was*, *were* and *-ed*, and instances removed from the count; to count these auxiliaries separately would double the count of verb phrases in which they appear. On manual inspection, 96% of automatically-extracted instances were judged to be actual present tense verbs. Results may be accurate enough to make this final step unnecessary.

Conventions. Instances of 31 kinds of conventions were identified manually. Extraction of this feature cannot be automated without substantial work to discern complex, reoccurring patterns in text which deviate from regular grammatic usage. Examples include: 1) news items beginning with location and source in parentheses: *SALEM (UP)* ...; 2) subsequent reference to a term by initials after its introduction. The short form is often given in parentheses immediately following the term's first use: *general assistance (GA) ... GA is guaranteed* ...; 3) nicknames: *Jimmy [Monk] Allegretti*; 4) various ways in which sports statistics are reported inline in text: *their total passing yardage in three games, 447 on 30 completions in 56 attempts, is only 22 yards short* ....

Citations. Instances of [{(and)}] bracket pairs containing zero or one space or a numeric field adjacent to the closing bracket were extracted automatically from text files. The results were checked manually to identify actual citations. While the simple pattern used ensured that recall was as close to complete as possible, 96% of automatically extracted instances were not citations. In retrospect, a few more discerning patterns could have been used, but for reporting accuracy we show the results obtained by a recall-maximizing pattern.

Title, Headings. Lines all in uppercase or less than 10 characters long were automatically extracted from text files. A check against longer lines was unnecessary in Brown because all titles in that corpus are set in capitals. Due to this special circumstance, recall was 100%.

Introduction. Lines containing the word *introduction*, *preface*, *prologue*, *preamble* or *foreword* were extracted automatically from text files. On inspection, 15% were judged to be overt introductory sections.

Passives. The .tag file was searched for tag sequences that may be given to passive constructions: /MD + "have" + "be" + /VBN, (/VBZ || /VBP) + "being" + /VBN, /VBD + "been" + /VBN, /MD + "be" + /VBN, (/VBZ || /VBP) + /VBN, ("am" || "are" || "is" || "was" || "were")/VBD + VBN. These patterns match such verb phrases as, respectively, *will have been carried*, *is being carried*, *has been carried*, *will be carried*, *is carried*, *am carried*. The complex structure of passive phrases allowed fully automatic handling of this feature: all automatically-extracted instances were true passive verbs.

Examples. The Unix utility *grep* was used to extract automatically lines containing instances of *example, typical, i.e., e.g., i. e., e. g., instance, illustration*. 72% of the extracted lines referred to explicit extended examples. We only counted candidates that referred to extended textual elaborations in the text, not illustrations or figures, and ones that were not full enumerations, that is, cases when all members of a group were explicitly listed.

Interrogatives. Instances of sentences containing a question mark were extracted automatically, and the raw list was manually inspected to exclude quoted questions. Half (49.6%) of automatically extracted instances proved to be questions directed to the reader.

Better filtering based on the proximity of quotation marks to the question mark might allow this operation to be entirely automatic.

Binders. Binders are sequencers like *third* and words like *however*, which set out and qualify relationships between the concepts expressed in a text's clauses and sentences. Marcu's list of 461 discourse connectives enumerates the set fairly comprehensively. Instances of Marcu's connectives were extracted automatically. Precision and recall are necessarily 100%.

Word, **Sentence**, **Paragraph Length**. These were measured in the obvious way from data in texts' .sgml files. The Unix wc command was used to determine word count.

Unique Words, Vocabulary Words. Computed as the line count of texts' .freq and .voc files respectively.

Vocabulary Familiarity. Computed as the familiarity of entries in texts' .voc files.

As we proceeded, certain facts became apparent.

- Errors in the data produced by automated processing seem to appear equally in all texts regardless of their degree of technicality. This suggests that manual post-processing (performed for most of the 16 features under consideration) might be dispensed with in favour of an approximation to the exact measure of a feature.
- The manual pass over a feature's data turned up sets of typical and marginal instances. A record of decisions on marginal instances helped clarify the extent of the feature. For example, questions appear both in text directed to the reader and in dialogue. The latter are not included in counts for this feature.
- Some features occur in every text, others appear only occasionally. The added features (six bottom rows of Table 2) fall into the first class, as do binders, interrogatives, and verb voice and tense. Instances of citations, introductions, conventions, and titles and headings do not appear in many texts.
- The amount of manual post-processing required varied substantially among features. As already suggested, recognizing conventions and colloquialisms was by far the most time-consuming task. For automatable features, the time required by the manual pass was proportional to the length of the list of extracted feature values, adjusted for the inherent accuracy of the filter—passives are easier to recognize than present tense verbs.

5. Outcomes

Results (Table 3, top of next page) are indicative but not conclusive. Certain features correlate with technicality at a level that approaches but does not reach statistical significance. Word and sentence length, titles and headings, and passives correlate positively, colloquialisms and binders negatively. Vocabulary unfamiliarity has a slight positive relationship, unique and vocabulary words a slight negative one. Present tense, interrogatives and paragraph length appear unrelated to technicality.

Scarcity of data prevents conclusions about citations, examples and introductions but experience suggests that, in particular, the presence of citations in a text should be a strong indicator that it is technical. This is less certain for introductions and examples. Any measure of a text's technicality must take into account features that are always present and those that occur only occasionally, with the very rarity of the latter adding weight to their evidence.

CATEGORY	colloquialisms	present tense	conventions	citations	title, headings	introduction	passives	examples	interrogatives	binders	word length	sent length	para length	unique words	vocab words	vocab faml	technicality
learned	0	49	2.5	1.36	5.7	0.1	20.6	1.76	1.19	84	6.17	23.4	122	708	600	2889	4.75
religion	1	60	1.0	0.00	4.7	0.0	16.0	0.71	4.06	83	5.88	24.3	118	776	643	2797	3.61
organizational	3	58	1.5	0.00	14.5	0.0	23.7	0.83	0.80	79	6.25	22.9	74	719	568	2332	3.61
press: editorial	7	68	1.0	0.00	13.4	0.0	16.5	0.85	4.59	95	5.96	18.8	60	904	712	2685	3.25
skills and hobbies	5	66	1.3	0.03	7.1	0.0	19.6	1.17	3.97	96	5.96	19.6	71	799	650	2940	3.22
press: reportage	6	52	2.2	0.00	8.0	0.0	15.4	0.36	0.52	92	6.02	19.9	46	895	630	2532	3.06
press: reviews	8	42	1.3	0.00	4.5	0.0	15.0	0.82	1.41	91	6.04	21.3	59	990	769	3443	2.75
popular lore	4	54	1.0	0.00	1.6	0.0	13.7	1.00	2.46	91	5.92	22.1	100	818	680	2994	2.55
belles lettres etc	7	55	0.0	0.48	0.9	0.0	12.9	1.13	2.31	86	5.94	23.8	133	826	691	2969	2.01
science fiction	5	60	1.0	0.00	4.5	0.0	6.7	0.33	2.33	123	5.80	13.8	37	826	705	2895	1.56
humor	9	55	1.5	0.00	1.0	0.0	7.2	0.67	3.22	100	5.80	19.4	102	899	761	3101	1.23
general fiction	6	58	1.0	0.00	2.7	0.0	3.5	0.03	4.00	117	5.56	15.8	76	802	693	3143	1.23
romance and love	8	62	1.0	0.00	2.6	0.0	3.2	0.17	3.21	122	5.54	14.9	58	776	668	2788	1.16
adventure & western	9	51	1.0	0.00	2.7	0.0	2.8	0.00	1.10	127	5.56	13.6	56	795	693	3176	1.15
mystery & detective	13	61	2.0	0.00	3.7	0.0	4.3	0.13	3.42	130	5.52	13.1	48	761	656	2589	1.13

Table 3: Indicator Feature Counts and Measures, Per Brown Corpus Category

Conventions are a special case for another reason. While certain conventions should be precise indicators of genre, it is premature to envision detecting them automatically. Work must first be done to identify and formalise conventions in general use.

6. Future Work and Conclusion

Feature value extraction should be run on a variety of other texts to see if the data it gathers confirm or correct the results obtained to date. Additional data could allow sparsely occurring features to be better analysed.

We also plan to see whether the raw data produced by the automated phase can serve successfully as an approximation without human editing. The phenomena involved here are such that no single feature or small group of features are likely to *guarantee* a particular degree of technicality for a text. The relationships are not strong enough, and we have learned that human intuitions about which features are good predictors are not always borne out by the data. On the other hand, the data consistently exhibit a vexing degree of correlation—not conclusive, but definitely not random. In fact, the notion of text technicality seems to be taken for granted: if one thing is clear in our results, it is the fact that technicality is a relative and rather subjective notion. It may be worthwhile to revisit the original task of feature identification, this time with the emphasis on measurement automation rather than plausibility. There may be other features that could be automated equally easily, and the current or an enlarged set may clearly justify a formula or computational mechanism that produces a technicality rating in accordance with that assigned by a human rater.

References

__ (1996). SAS/STAT Software: Changes and Enhancements through Release 6.11. SAS Institute Inc.

ALBERT, J.H. & S. CHIB (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of American Statistical Association*, 88, pp.669-679.

APTÉ, C., F. DAMERAU & S. WEISS (1994). Automated Learning of Decision Rules for Text Categorization. *ACM Transactions on Information Systems*, 12 (3), pp.233-251.

BIBER, D. (1988). *Variation Across Speech and Writing*. University Press, Cambridge.

BRILL, E. (1992). A Simple Rule-based Part of Speech Tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*, ACL, pp.152-155.

CARLETTA, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2), pp.249-254.

CHANDRASEKAR, R. & B. SRINIVAS (1997). Using Syntactic Information in Document Filtering: A Comparative Study of Part-of-speech Tagging and Supertagging. *Proceedings of the RIAO-97 Conference*, pp.531-545.

COHEN, W. (1999). Context-sensitive Learning Methods for Text Categorization. *ACM Transactions on Information Systems*, 17 (2), pp.141-173.

COPECK, T., K. BARKER, S. DELISLE & S. SZPAKOWICZ (1999), More Alike than not—An Analysis of Word Frequencies in Four General-purpose Text Corpora. *Proceedings of the 1999 Pacific Association for Computational Linguistics Conference (PACLING 99)*, pp.282-287.

COPECK, T., K. BARKER, S. DELISLE, J.-F. DELANNOY & S. SZPAKOWICZ (1997), What is Technical Text? *Language Sciences*, 19 (4), pp.391-424.

FRANCIS, W.N. & H. KUCERA (1982). *Frequency Analysis of English Usage*. Houghton Mifflin, Boston.

FROISSART, C. & G. LALLICH-BOIDIN (1999). Le document technique : unicité et pluralité. *Actes du 2ème Colloque du Chapitre Français de l'ISKO — L'indexation à l'ère d'internet*, pp.18.1-18.9.

IKEDA, T., A. OKUMURA & K. MURAKI (1998). Information Classification and Navigation based on 5W1H of the Target Information", *Proceedings of the COLING-ACL'98 Conference*, pp.571-577.

ILLOUZ, G., B. HABERT, H. FOLCH, S. HEIDEN, S. FLEURY, P. LAFON & S. PRÉVOST (2000). TyPTex: Generic Features for Text Profiler. *Proceedings of the RIAO-2000 Conference*, pp.1526-1540.

JOACHIMS, T. (1999). Transducive Inference for Text Classification Using Support Vector Machines. *Proceedings of the 1999 International Conference on Machine Learning (ICML)*.

JUNKER, M. & A. ABECKER (1997). Exploiting Thesaurus Knowledge in Rule Induction for Text Classification. *Proceedings of the RANLP-97 Conference*, pp.202-207.

KARLGREN, J. & D. CUTTING (1994). Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. *Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING-94)*, pp.1071-1075.

KILGARRIFF, A. (1995). Review of Dimensions of Register Variation: A Cross-Linguistic Comparison. *Journal of Natural Language Engineering*, 1 (4), pp.396-399.

LARKEY, L. & B. CROFT (1996). Combining Classifiers in Text Categorization. *Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval (SIGIR 96)*, pp.289-297.

LEWIS, D.D. & R.M. TONG (1992). Text Filtering in MUC-3 and MUC-4. *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, pp.51-66.

LIDDY, E., W. PAIK & E. YU (1994). Text Categorization for Multiple Users Based on Semantic Features From a Machine-Readable Dictionary. *ACM Transactions on Information Systems*, 12 (3), pp.278-295.

MARCU, D. (1998). A Surface-Based Approach to Identifying Discourse Markers and Elementary Textual Units in Unrestricted Texts. *Proceedings of the COLING-ACL'98 Workshop on Discourse Relations and Discourse Markers*, pp.1-7.

MILLER, G. (1990). WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4).

NISKE, H. (1998). Text Linguistic Models for the Study of Simultaneous Interpreting. Licentiate, Stockholm University.

PAPKA, R. & J. ALLEN (1998). Document Classification Using Multiword Features. *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM)*.

SCOTT, S. & S. MATWIN (1998). Text Classification Using WordNet Hypernyms. *Proceedings of the COLING-ACL'98 Workshop on Usage of WordNet in Natural Language Processing Systems*, pp.45-52.

SIEGEL, S. & N. J. CASTELLAN (1988) *Nonparametric Statistics for the Behavioral Sciences* (2nd ed.). New York: McGraw-Hill.

SINCLAIR, J. & J. BALL (1996). Preliminary Recommendations on Text Typology. Expert Advisory Group on Language Engineering Standards (EAGLES). #EAG-TCWG-TTYP/P.

TROSBORG, A. (1999). *Text Typology and Translation*. ed. John Benjamins, Amsterdam.

WILBUR, W. (1996). An Analysis of Statistical Term Strength and its Use in the Indexing and Retrieval of Molecular-Biology Texts. *Computers In Biology And Medicine*, 26 (3), pp.209-222.

YANG, Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval Journal*, 1 (1-2), pp.69-90.