# West Group at CLEF2000: Non-English Monolingual Retrieval

Isabelle Moulinier, J. Andrew McCulloh, Elizabeth Lund
West Group
610 Opperman Drive
Eagan, MN 55123
USA
Isabelle.Moulinier@westgroup.com

West Group participated in the non-English monolingual retrieval task for French and German. Our primary interest was to investigate whether retrieval of German or French documents was any different from the retrieval of English documents. We focused on two aspects: stemming for both languages and compound breaking for German, and studied several query formulations to take advantage of compounds. Our results suggest that German retrieval is indeed different from English or French retrieval, inasmuch as breaking compounds can significantly improve performance.

## Introduction

West Group's first attempt at non-English monolingual retrieval was through its participation in the Amaryllis-2 campaign. Our findings during that campaign were that there was little difference between French and English retrieval, once the inflectional nature of French was handled through stemming or morphological analysis. For CLEF-2000, our goal for French document retrieval was to investigate the impact of the stemming method. We compare performing no stemming, stemming using an inflectional morphological analyzer, and stemming used a rule-based algorithm similar to Porter's English stemmer.

Our main focus, however, was German document retrieval. German introduced a new dimension to our previous work: compound terms. We set up our experiments to assess whether we could ignore compound terms, i.e. handle German retrieval like we handled French or English retrieval, or whether we could leverage from the existence of compounds.

For both our French and German experiments, we relied on a slightly altered version of the WIN engine, West Group's implementation of the inference network retrieval model [Tur90]. We used third-party stemmers to handle non-English languages.

In the following, we briefly describe the WIN engine and its adaptation to non-English languages. Next, we describe our variants for German document retrieval. The section following that describes experiments with stemming for French monolingual retrieval.

## General System Description

The WIN system is a full-text natural language search engine, and corresponds to West Group's implementation of the inference network retrieval model. While based on the same retrieval model as the INQUERY system [BCC93], WIN has evolved separately and focused on the retrieval of legal material in large collections in a commercial environment that supports both Boolean and natural language searches [Tur94].

The WIN engine supports three types of document scoring: the document as a whole is scored; each paragraph is scored and the score of the document becomes the best paragraph score; the document score and the best paragraph score are combined. We used the following scoring techniques:

- German: document score based on whole document
- French: document score based on combination of whole document and best paragraph.

We indexed non-English collections using a slightly modified WIN for each language:

- German:
  - We used a third-party stemmer based on a morphological analyzer. One of the features was compound decomposition. Forcing decomposition or not was a parameter in our experiments.
  - We indexed both German collections as one single retrieval collection. We did not investigate merging retrieved sets.
- French
  - We added a tokenization rule to handle elision.
  - We used two kinds of stemmers: a third-party stemmer based on a morphological analyzer, and a rule-based stemmer (*a la* Porter) from the Muscat project.

A WIN query consists of concepts extracted from natural language text. Normal WIN query processing eliminates stopwords, noise phrases (or introductory phrases) and recognizes phrases or other important concepts for special handling. Many of the concepts ordinarily recognized by WIN are specific to both English documents and the legal domain. To perform these tasks, WIN relies on various resources: a stopword list, a list of introductory phrases ("Find cases about…", "A relevant document describes…") , a dictionary of (legal) phrases.

Query processing for French was similar to English query processing. We used a stopword list of 1745 terms (highly frequent terms, and noise terms like adverbs). Using the TREC-6, 7 and 8 topics, we refined the list of introductory patterns we created for Amaryllis-2. In the end, there were 160 patterns (a pattern is a regular expression that handles case variants and some spelling errors). We did not use phrase identification for lack of a general French phrase dictionary.

We investigated several options to structuring German queries, decomposing or not decomposing compounds. This specific processing is described below. We used a stopword list of 333 terms. Using the TREC-6, 7 and 8 topics, we derived a set of introductory patterns for German. There were 11 regular expressions, summarizing over 200 noise phrases. We did not perform phrase identification through a dictionary. However German compounds have been treated as "natural phrases" in some of our runs.

Finally, we extracted concepts from the full topics. However, we gave more weight to concepts appearing in the Title or Description fields than concepts extracted from the Narrative field. Following West's participation at TREC3 [TTYF95], we assigned a weight of 4 to concepts extracted from the Title field, while concepts originating from the Description and Narrative fields were given a weight of 2 and 1, respectively.

## German monolingual retrieval experiments and results

Our experiments with monolingual German retrieval focused on query processing and compound decomposition. Our submitted runs rely on decomposing compounds, but we also experimented with no decomposition, and no stemming at all. Indexing followed the choice made for query processing. For instance, when no decomposition was performed for query terms, parts of compounds were not indexed.

When dealing with breaking compound terms, we faced the choice of considering a compound term as a single concept in our WIN query, or treating the compound as several concepts (as many concepts as there were parts in the compound). The submitted run WESTgg1 considers that a compound corresponds to several concepts; the run WESTgg2 handles a compound as a single concept.

When faced with a compound Energiequellen, the structured query in WESTgg1 introduces 2 concepts, Energie and Quelle; the structured query in WESTgg2 introduces 1 concept, #PHRASE(Energie Quelle). The #PHRASE operator is a soft phrase, i.e. the component terms must appear with 3 words of one another. The score of the #PHRASE concept in our experiment was set to be the maximum score of the soft phrase itself or of its components.

Table 1 summarizes the results of our two official runs as well as the results of the runs Nostem where no stemming was used and Nobreak where stemming but no decomposition was used.

| | | | Performance of individual queries | | | | |
|---|---|---|---|---|---|---|---|
| Run | Avg. Prec. | R-Prec. | Best | Above | Median | Below | Worst |
| WESTgg1 | 0.3840 | 0.3706 | 3 | 21 | 3 | 9 | 1 |
| WESTgg2 | 0.3779 | 0.3628 | 3 | 18 | 6 | 9 | 1 |
| Nostem | 0.2986 | 0.3080 | 0 | 15 | 1 | 19 | 2 |
| Nobreak | 0.2989 | 0.3141 | 0 | 18 | 1 | 15 | 3 |

**Table 1: Summary of individual run performance on the 37 German topics with relevant documents.**

The results reported in Table 1 support the hypothesis that German document retrieval differs from English document retrieval. Treating compound words as forms of phrases improves the performance of the German retrieval system. Indeed, searching with compound stems did not perform better than searching with no stemming.

We expected a greater difference between our two submitted runs. WESTgg1 allows compound terms to contribute more to the score of a document, while WESTgg2 gives the same contribution to compound and non-compound terms. The contribution of a compound term in WESTgg1 is weighted by the number

of parts in the compound, so one would expect its occurrence in a document to significantly alter a document score.

After reviewing the individual queries, we noticed the following behavior. First, for those queries where both the compounds and their parts had an average frequency, neither particularly common nor particularly rare, the two runs behaved similarly. Then, the parts helped locate documents, but did not add to or draw away from the document relevance score. Second, for those queries where the compound itself is above average, but the individual parts are average, or even fairly common, then the weighted contribution provided in WESTgg1 performed better. Third, for those queries where at least one part of a compound was very common, the high occurrence of that part degraded the weighting scheme of WESTgg1, thus the single concept construct of WESTgg2 provided a more representative score.

Also, compound handling in WESTgg1 as well as WESTgg2 is only as influential as there are compounds in the query. In the 40 German topics, roughly 16% of the query terms are compound terms. In addition, it should noted that we indexed the individual parts of compounds. As a result, a simple query term may also match the part of a compound in a document.

## French monolingual retrieval experiments and results

The goal of our experiments with French document retrieval was to assess the difference between stemming algorithms. Our motivation was to further investigate the particularity of French compared to English. [Hul96] reported results on various kinds of stemmers for English document retrieval. So far, we have studied two types of stemmers (out of the 5 types in [Hul96]) as well as no stemming at all:

- a stemmer based on an inflectional morphological analyzer, e.g. it conflates verb forms to the infinitive of the verb, noun forms to the singular noun, adjectives to the masculine singular form. This stemmer is based on a lexicon.
- a rule-based stemmer "a la Porter" that approximates mainly inflectional rules, but also provides a limited set of derivational rules based on suffix stripping, e.g. it strips suffixes like –able or –isme.

Our runs also took advantage of the multiple TEXT elements in a document. We considered those elements to mark paragraph boundaries and used this information for document scoring and ranking.

Our submitted run, WESTff, used the inflectional stemmer. Table 2 summarizes the performance of runs using the inflectional stemmer, the Porter stemmer and no stemmer at all. We also ran experiments when a document was either scored as a whole or as its best paragraph. Those runs are not reported here, as they did not perform as well as the combined score.

| | | | Performance of individual queries | | | | |
|---|---|---|---|---|---|---|---|
| Run | Avg. Prec. | R-Prec. | Best | Above | Median | Below | Worst |
| WESTff | 0.4903 | 0.4371 | 11 | 9 | 7 | 7 | 0 |
| Porter | 0.4680 | 0.4297 | 6[1] | 14 | 1 | 13 | 0 |
| Nostem | 0.4526 | 0.4210 | 7[1] | 8 | 0 | 19 | 0 |

**Table 2: Summary of individual run performance on the 34 French topics with relevant documents.**

While we usually consider not stemming as a baseline, our tests showed that no stemming performed better on several topics. In those instances, we found that the Porter stemmer was too aggressive and stemmed important query terms to very common forms. For instance, parti was stemmed to part, directive to direct and français to franc. The inflectional stemmer did exactly what it was supposed to do, e.g. stem française to français. However, certain stems were very common, while their raw form was less common. Phrase identification, e.g. académie française, and monnaie européenne, may likely improve performance, as it has proven to be beneficial for the English version of the WIN search engine.

In addition, the inflectional stemmer is only as good as its lexicon. We found a couple of queries where the Porter stemmer performed better because important query terms were not in the lexicon.

While our analysis is only partial at this time, it appears that our French stemming results follow the patterns exhibited by [Hul96] for English stemming, except that inflectional stemming seems slightly superior. We do not know yet whether this is a particularity of the French language or of this particular collection and set of topics.

[1] For some queries, our runs achieved an average precision that was better than the best average precision reported at CLEF.

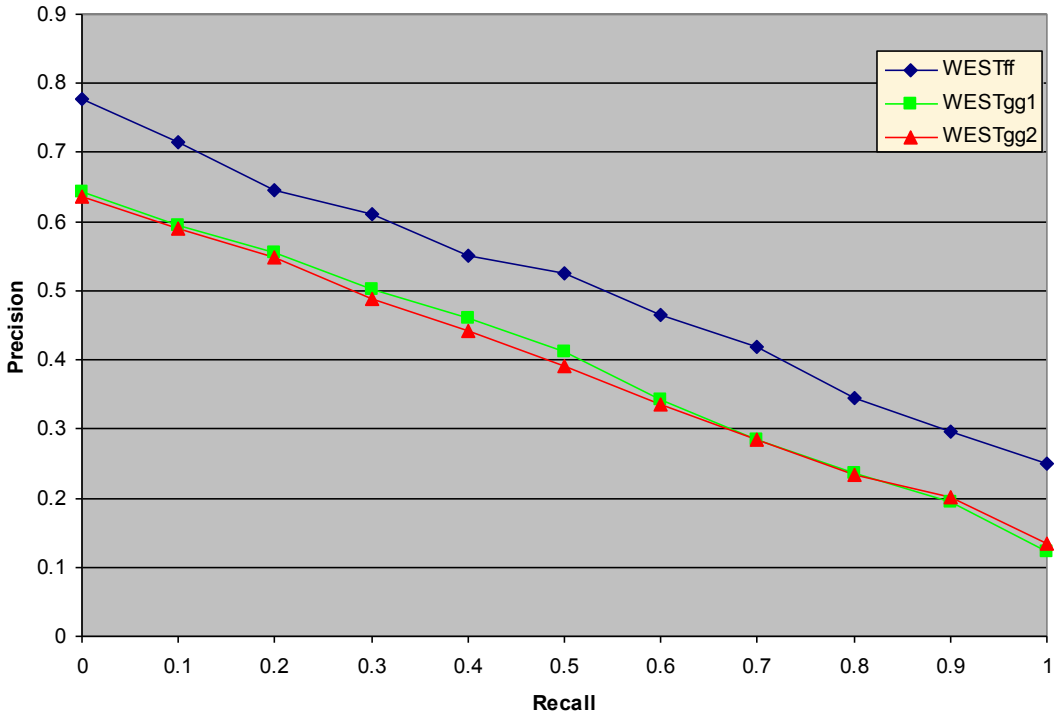


**Figure 1: Recall-precision curves for our submitted runs in non-English monolingual document retrieval**

## Summary

The WIN retrieval system achieved good performance for both German and French document retrieval without any major modification being made to its retrieval engine. On the one hand, we showed that German document retrieval required specific handling because of the use of compound words in the language. Our results showed that decomposing compounds during indexing and query processing enhanced the capabilities of our system. Our French experiments, on the other hand, did not uncover any striking difference between French and English retrieval, except a preference towards the use of an inflectional stemmer.

## References

[CCB92] W.B. Croft, J. Callan and J. Broglio. The INQUERY retrieval system. *In Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, Spain, 1992

[Hul96] D. A. Hull. Stemming Algorithms: A Case Study for Detailed Evaluation. *In Journal of The American Society For Information Science*, 47(1): 70-84,1996.

[TTYF95] P. Thompson, H. Turtle, B. Yang and J. Flood, "TREC-3 Ad Hoc Retrieval and Routing Experiments using the WIN System," in *Overview of the 3rd Text Retrieval Conference (TREC-3),* NIST Special Publication 500-225, Gaithersburg, MD, April 1995.

[Tur90] H. Turtle. *Inference Networks for Document Retrieval*. PhD Thesis, Computer Science Department, University of Massassuchets, Amherst, 1990.

[Tur94] H. Turtle. Natural language vs. Boolean query evaluation : a comparison of retrieval performance. In *Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval*, Dublin, 1994