Statistical Stemmers: A Reproducibility Study

Gianmaria Silvello, Riccardo Bucco, Giulio Busato, Giacomo Fornari, Andrea Langeli, Alberto Purpura, Giacomo Rocco, Alessandro Tezza, and Maristella Agosti

Department of Information Engineering, University of Padua, Italy
{name.surname}@unipd.it

Abstract. Statistical stemmers are important components of Information Retrieval (IR) systems, especially for text search over languages with few linguistic resources. In recent years, research on stemmers produced relevant results, especially in 2011 when three language-independent stemmers were published in relevant venues. In this paper, we describe our efforts for reproducing these three stemmers. We also share the code as open-source and an extended version of Terrier system integrating the developed stemmers.

1 Introduction

The research on stemmers has focused for a long time on the English language and successively on a subset of other, mostly European, languages. Hence, for English several highly-effective rule-based stemmers such as Porter [11], Krovetz [5] and Lovins [6] are commonly available. For highly studied languages such as German, French, Italian and Spanish the effective rule-based stemmers implemented by Snowball [1] are typically employed by IR systems. On the other hand, for languages as the Slavic or Asian ones, there are few linguistic resources and rule-based stemmers are less effective, if available at all. In these cases, statistical stemmers can play a key role, since being language-independent they can be employed without any prior knowledge of the language at hand. Despite their relevance, statistical stemmers are not commonly taken into account in baseline IR systems or considered in longitudinal studies in IR even when non-English corpora are considered [2,4].

In recent years, we have witnessed a new interest in research on statistical stemmers with a spike in 2011 when three new stemming algorithms were proposed by the same core subset of authors – i.e., Jiaul H. Paik and Swapan K. Parui – in relevant IR venues, namely: "A fast corpus-based stemmer" (FCB) published in the ACM TALIP [10], "A novel corpus-based stemming algorithm using co-occurrence statistics" (SNS) presented at SIGIR [9] and "GRAS: An effective and efficient stemming algorithm for information retrieval" published in the ACM TOIS [8]. More recently, these works have been reconsidered and discussed in a comprehensive survey about text stemming [13].

[1] http://snowballstem.org/

We decided to reproduce these three papers with the aim of making the stemmers they propose readily available to the research community such that they can be easily included in baseline systems and longitudinal studies. To this end, we also share an extended version of the Terrier system [7] where these stemmers have been integrated and are ready to be used in a typical IR experimental setting.

The paper is organized as follows: Section 2 introduces the stemmers and their main characteristics; Section 3 presents the main issues we faced when implementing the stemming algorithms; Section 4 reports the experimental results and the differences with the reference papers; finally, Section 5 discusses the pros and cons of the reproduced papers and what can be learned for the future.

2 Overview of the statistical stemmers

FCB. FCB is a statistical stemmer that relies on the frequency of suffixes of the terms in a language. FCB associates a frequency to each suffix appearing in a corpus of documents equal to the number of words that end with it. If the frequency of a suffix exceeds a certain threshold α, then that suffix is called a *potential* suffix. The algorithm then starts to group the terms in the collection into k-equivalence classes according to their prefix. It iteratively groups the terms with a prefix of length k or longer in common, at each step decreasing the minimum common prefix length until a pre-determined lower threshold. As suggested in the reference paper we set 5 as a starting prefix length and 2 as lower threshold. After the terms have been grouped into equivalence classes, the longest common prefix of each class is evaluated as a candidate stem for the class. To do this, FCB considers the size of the subsets of elements in the class that contains only terms whose suffixes, induced by the candidate stem, all belong to the potential suffixes set. Finally, we compute the ratio of the size of the largest of the aforementioned subsets – i.e., *potential-class* – and the size of the class to be evaluated. If this ratio exceeds a certain threshold δ, the longest prefix of that class is considered a valid stem for all of the terms. Otherwise, a better stem for the class is chosen amongst the values in it and the evaluation process is repeated by iteratively extracting all the subsets of terms that have a common valid stem in the class.

SNS. The goal of SNS is to group words that are morphologically related by computing their co-occurrence in the corpus. The starting hypothesis is that a document relevant to a given topic will probably contain many words relevant to the topic itself. SNS is composed of three main steps: (i) the computation of the co-occurrence strength of word pairs; (ii) the re-calculation of the strengths; and (iii) the clustering of the words. Once the co-occurrence strengths (CO) are computed, they are represented in a weighted graph where words, say $w1$ and $w2$, are nodes connected by an edge if at least one of these two conditions holds: (i) $CO(w1, w2) > 0$; and (ii) if "they have a common prefix of a given length (l_1) along with the suffixes (after truncating the longest common prefix) which

are the residues (the ends after removal of longest common prefix) of more than one pair of co-occurring words with long common prefix (length larger than 5 which is the second static parameter, l_2)" [9].

The second step performs the re-calculation of co-occurrence strength between word pairs. The strength assigned to a word pair $(w1, w2)$ is proportional to the number of other words in the corpus that co-occur with both $w1$ and $w2$. Afterwards, the co-occurring words are clustered together. To do this SNS identifies the strong edges; an edge (u, v) is defined as "strong" if two conditions hold: (i) (u, v) has the highest weight w.r.t all the edges insisting on u; and, (ii) (u, v) has the highest weight w.r.t all the edges insisting on v. Lastly, the non-strong edges are removed from the graph and the remaining connected components of the graph represent the morphologically related groups.

GRAS. GRAS is a stemmer conceived for highly inflectional languages (e.g. Hungarian) where words are formed from the root by a process of suffixation. Hence, the role of suffixes is central for this algorithm. GRAS can be described as a sequence of five main steps.

In the first step, GRAS identifies the word partitions sharing a l-long prefix; l is set to be the average word length for the given language. The second step determines the common suffixes of the words sharing a prefix and it checks if there exist other word pairs with a common prefix followed by the same identified suffix. Two suffixes are considered a candidate pair if they are shared "frequently enough" by word pairs in the lexicon; the suffix frequency threshold is defined by a parameter α. In the third step, GRAS creates a graph where the identified words are mapped to nodes which are connected by an edge if the words are morphologically related – i.e., they share a non-empty prefix and the suffix pairs that remain after the removal of the common prefix are candidate pairs identified in the second phase. In the fourth phase, *pivot* nodes – words with a large number of edges — are identified. In the fifth step, equivalence classes of words are created. A word is put in the same class as the pivot to which it is connected if it has a *cohesion* of at least δ. The cohesion value determines the likelihood that two words – the candidate word and the pivot – are morphologically related.

3 Realization of the stemmers

All the described stemmers have been implemented in Java and integrated into Terrier v4.1. [2] The input of the stemmer is composed of the lexicon and the inverted index created by Terrier and the output is a text file (i.e., the `lookup table`) containing the words in the lexicon and their stems. We extended Terrier v4.1 in order to use the statistical stemmers we realized.

FCB. Since FCB relies on the frequency of suffixes in a given corpus, we developed a suffix extraction algorithm. In the reference paper there is no description of the suffix extraction process, therefore we decided to extract all the suffixes

[2] `http://github.com/giansilv/statisticalStemmers/`

in a given corpus, without considering the inclusion relations between them. For example, if we assume that for English the suffix "ing" is a possible suffix, we extract and compute the frequency of the suffixes "g", "ng" and "ing". Afterwards, we employed the "frequency-based filtering strategy" described in the reference paper to select the most frequent suffixes.

With regards to the implementation of the core of the stemming algorithm, we made an assumption about the evaluation of the k-equivalence classes. A *potential-class* is defined in the reference paper as the largest subset of words with a common prefix R ending with frequent suffixes induced by R. However, we realized two versions of the algorithm: FCB v.1 and FCB v.2. FCB v.1 considers the prefix R to compute the suffixes, whereas FCB v.2 ignores it. This is a crucial part of the algorithm and more details about how it has been realized would have been important for reproducibility purposes. FCB v.1 considers the strings composed of all the characters that follow the common prefix R as suffixes of the terms in a k-equivalence class and then compares them against the set of frequent suffixes extracted from the collection prior to the execution of the algorithm. FCB v.2 on the other hand considers the whole terms in a k-equivalence class and qualifies them as ending with a frequent suffix if they end with any of the frequent suffixes mentioned before. Therefore, in this case, we allow the presence of a few characters between R and the beginning of the suffix. We noticed a great improvement in the performances of the algorithm by using the latter approach, especially for languages with greater inflection such as Hungarian.

Finally, the reference paper does not describe how to deal with singleton classes; singleton classes have as longest common prefix the whole term that belongs to the class, therefore, in this case, the induced suffix is always empty. This implies that many terms with a unique prefix, but ending with a frequent suffix are not stemmed.

SNS. Our implementation of SNS is composed of six main steps. *Step 1.* We extract the data from the lexicon and the inverted index and contextually we discard the terms whose first character is a digit or that are shorter than l_1 ($l_1 = 3$ in the reference paper) in order to avoid useless computations. *Step 2.* We compute the co-occurrence between two words if they have a common prefix with length greater than or equal to l_1; if their strength is not zero, then we check if their common prefix is greater than or equal to l_2 ($l_2 = 5$ in the reference paper) and we store this information to be used in later phases. *Step 3.* We create a weighted graph where words are nodes and the edges are weighted by their co-occurrence strength. *Step 4.* We update the edge strength by re-calculating the co-occurrence of terms. *Pass 5.* We remove the non-strong edges. *Step 6.* We find the connected components of the graph. Each stem is generated by finding the longest common prefix amongst the connected words; this is an assumption we made since in the reference paper this phase is not described and the proposed algorithm stops right after the creation of word clusters.

GRAS. In the first step of GRAS implementation, we process the lexicon by creating partitions of words sharing a common prefix of length l. This step

reduces the size of the lexicon at hand and reduces the running time of the algorithm. In the second step, for each common prefix class, we individuate the α-frequent suffixes and we store them along with their frequencies. In the third step, we build the graph (one for each common prefix class) and we calculate the cohesion for each pair of connected words.

As we discuss below, in the reference paper there are a few moot points that may create some problems from the reproducibility viewpoint. First-of-all, the l parameter is set to be "the average word length for the language concerned", but no further details are given. We chose to calculate the average length of the words in the lexicon at hand; we would have liked to have the actual l parameter employed in the reference paper since this parameter has a high influence on the stemmer and a small difference here can be sizable performance-wise.

4 Reproduction of the results

The results for FCB, SNS and GRAS have been reproduced for the CLEF and TREC collections employed in the reference papers; more details are reported below. CLEF collections have been downloaded from DIRECT [3] and TREC collections from the TREC Website. [4] In the reference papers also the Marathi and Bengali test collections from FIRE are employed, but these collections are not currently available in the FIRE Website. [5] Nevertheless, the CLEF and TREC collections form a solid comparative testbed for assessing our reproducibility effort. All the experiments were conducted using Terrier v4.1 (with the IFB2 model) as a baseline system. All the stemmers were tested after a stopwords removal phase; for the English language we adopted the stoplist provided by Terrier and for the other languages those provided by Jacques Savoy. [6] In the following we show the results obtained by the original algorithms and the reproduced one and we report their absolute differences.

FCB. The algorithm was evaluated on the CLEF 2006-2007 collection for the Hungarian language (98 topics) and on the Wall Street Journal sub-corpus of the TIPSTER collection (topics 1-200). The experiments are divided into two sets, one where the queries are composed of the topic title (T) and the other where the queries are composed of title and description (TD). The version of Terrier employed in the reference paper is not specified. The experiments were performed by removing the stopwords and numbers from the lexicon before the execution of the stemming algorithm. We tested both the versions of the algorithm we realized for different values of δ. In the following tables, we report the results achieved with the best tested δ value.

In Table 1 we report the results obtained for English. In this case, we employed FCB v.1 and even though there are sizable differences between our implementation and the original stemmer, our results are consistent with the original

[3] `http://direct.dei.unipd.it/`
[4] `http://trec.nist.gov/`
[5] `http://fire.irsi.res.in/fire/static/data/`
[6] `http://members.unine.ch/jacques.savoy/clef/`

Table 1: English WSJ TREC collection. Original and reproduced (FCB v.1) results for $\delta = 0.6$, the reference paper reports only 3 digits after the decimal point. Differences greater than 0.0100 are reported in bold.

		Original			Reproduced			Difference		
		MAP	RPrec	P@10	MAP	RPrec	P@10	MAP	RPrec	P@10
T	No Stem	0.225	0.267	0.399	0.2250	0.2674	0.3990	0.000	0.000	0.000
	FCB	0.258	0.289	0.437	0.2399	0.2791	0.4020	**-0.018**	**-0.010**	**-0.035**
	Porter	0.261	0.296	0.432	0.2621	0.2971	0.4362	+0.001	+0.001	+0.004
TD	No Stem	0.272	0.312	0.477	0.2722	0.3125	0.4765	0.000	+0.000	0.000
	FCB	0.295	0.331	0.493	0.2811	0.3181	0.4715	**-0.014**	**-0.013**	**-0.0215**
	Porter	0.294	0.325	0.477	0.2958	0.3262	0.4800	+0.002	+0.001	+0.0030

ones since FCB is better than no stemmer and worse than Porter; this is reasonable for a language with good linguistic resources.

Table 2: CLEF 2006-2007 Hungarian collection. Original and reproduced (FCB v.2) results for $\delta = 0.5$, the reference paper reports only 3 digits after the decimal point. Differences greater than 0.0100 are reported in bold.

		Original			Reproduced			Difference		
		MAP	RPrec	P@10	MAP	RPrec	P@10	MAP	RPrec	P@10
T	No Stem	0.185	0.199	0.258	0.1830	0.1956	0.2547	-0.0020	-0.0034	-0.0033
	FCB	0.293	0.315	0.353	0.2863	0.2942	0.3284	-0.0067	**-0.0208**	**-0.0246**
	RB	0.267	0.280	0.343	0.2610	0.2737	0.3245	-0.0060	-0.0063	**-0.0185**
TD	No Stem	0.239	0.252	0.314	0.2375	0.2528	0.3133	-0.0015	+0.0008	-0.0007
	FCB	0.341	0.352	0.390	0.3355	0.3263	0.3949	-0.0055	**-0.0257**	+0.0049
	RB	0.335	0.340	0.389	0.3347	0.3358	0.4102	-0.0020	**-0.0532**	**+0.0212**

In Table 2, we report the results obtained for Hungarian by employing FCB v.2, which turns out to be quite stable to the variations of δ. We can see that in this case the difference between the original algorithm and FCB v.2 is smaller than for English. There can be a sizable difference between our Terrier setup and the one of the reference paper, since also for the no stemmer and the rule-based stemmer (RB) cases we register a difference comparable to the one we get for FCB.

In order to enable the comparison of performances between FCB and the other stemmers we reproduced, we tested FCB also on the Bulgarian and Czech CLEF collections that are considered by both SNS and GRAS; the results are reported in Table 3.

SNS. SNS was evaluated on three CLEF collections – i.e. the 2006-2007 CLEF Bulgarian and Hungarian collections and the CLEF 2007 Czech collection – and one TREC collection – i.e. the corpus is the TIPSTER Disk 4&5 minus

Table 3: FCB v.2 results for the Bulgarian (2006-2007) and Czech (2007) CLEF collections with $\delta = 0.5$.

		Bulgarian			Czech		
		MAP	RPrec	P@10	MAP	RPrec	P@10
T	No Stem	0.165	0.191	0.220	0.220	0.248	0.244
	FCB	0.239	0.270	0.293	0.306	0.306	0.314
TD	No Stem	0.203	0.230	0.257	0.238	0.261	0.268
	FCB	0.276	0.301	0.333	0.338	0.318	0.348

the congressional record and the federal register and the TREC6, TREC7 and TREC8 topics. The queries used for the experiments are composed of the title and description fields of the topics. The results in the reference paper were obtained using Terrier, the version of which is not specified.

In Table 4, we report the results for Bulgarian for a system employing no stemmer, a rule-based stemmer and SNS. The rule-based stemmer employed in the reference paper for Bulgarian is not further specified. There are at least three rule-based stemmers that can be used: an aggressive one, a light one using transliterated terms and one which is the same as the light stemmer except that it processes documents in Cyrillic. The closest performance value to the original one is obtained with the third stemmer and it is the one we report in the result table.

Table 4: CLEF 2006-2007 Bulgarian test collection. Original and reproduced results for the SNS stemmer and difference between the reproduced stemmer and the original one. Differences greater than 0.0100 are reported in bold.

	Original			Reproduced			Difference		
	NO	RB	SNS	NO	RB	SNS	NO	RB	SNS
MAP	0.2166	0.2794	0.3256	0.2038	0.2786	0.2980	**-0.0128**	-0.0008	**-0.0276**
RPrec	0.2293	0.2930	0.3289	0.2291	0.3033	0.3253	-0.0002	**+0.0103**	-0.0036
P@10	0.2570	0.3270	0.3520	0.2580	0.3410	0.3540	+0.0010	**+0.0140**	+0.0020

From Table 4 we can see that there is a sizable difference in terms of MAP between the systems not employing any stemmer; this difference is increased when we consider the MAP values for the systems employing SNS. As we can see for RPrec and P@10 the difference is quite small and it does not affect the reproduced results in an appreciable way. Another difference can be seen for RPrec and P@10 when the rule-based stemmer is employed.

In general, we see that SNS improves the baseline systems (no stemmer and rule-based stemmer) both in the reference paper and in the reproduced version, even though in the reproduced case the improvement is less marked, especially in terms of MAP. The problem with the reproducibility of this stemmer for Bulgarian seems to be related to the starting difference between the baseline

systems (i.e. no stemmer) rather than to the specific implementation of SNS. This fact can be further assessed by considering the results for the other test collections.

For Czech, the authors claim to use the stemmer defined in [3], which actually presents two rule-based stemmers, one light and one aggressive. We tested both these stemmers and we found that the light one was used in the reference paper.

Table 5: CLEF 2007 Czech test collection. Original and reproduced results for the SNS stemmer and difference between the reproduced stemmer and the original one. Differences greater than 0.0100 are reported in bold.

	Original			Reproduced			Difference		
	NO	RB	SNS	NO	RB	SNS	NO	RB	SNS
MAP	0.2381	0.3409	0.3624	0.2382	0.3405	0.3569	+0.0001	-0.0004	-0.0055
RPrec	0.2611	0.3456	0.3441	0.2611	0.3456	0.3449	0.0000	0.0000	-0.0008
P@10	0.2680	0.3480	0.3700	0.2680	0.3480	0.3640	0.0000	0.0000	-0.0060

In Table 5 we can see that the results obtained by using no stemmer and the rule-based stemmer are reproduced with a very marginal error and thus we can state that the initial condition for the experiment with SNS is the same as in the reference paper. Also the results obtained with SNS are very close to the original ones with marginal differences for all the measures.

For Hungarian, we have a problem with the rule-based stemmer since the authors have specified that they use the stemmer defined in [12], where both a light and an aggressive stemmer are defined. We ran the experiments for these stemmers and we report the results obtained with the light stemmer that are closer to those in the reference paper.

Table 6: CLEF 2006-2007 Hungarian test collection. Original and reproduced results for the SNS stemmer and difference between the reproduced stemmer and the original one. Differences greater than 0.0100 are reported in bold.

	Original			Reproduced			Difference		
	NO	RB	SNS	NO	RB	SNS	NO	RB	SNS
MAP	0.2386	0.3132	0.3588	0.2375	0.3369	0.3583	-0.0011	**+0.0237**	-0.0005
RPrec	0.2518	0.3117	0.3585	0.2528	0.3459	0.3556	+0.0010	**+0.0342**	-0.0029
P@10	0.3143	0.3990	0.4224	0.3133	0.4153	0.4163	-0.0010	**+0.0163**	-0.0061

In Table 6 we report the original and the reproduced results for Hungarian, where we can see that the major differences concern the rule-based stemmer. Possibly, the authors employed a different rule-based stemmer than one of those defined in [12]. Nevertheless, the reproduced results for SNS (as well as for the

no stemmer system) are close to the original ones and this allows us to state that SNS has been correctly reproduced for Hungarian. We can also see that SNS is still slightly superior to the rule-based stemmer we employed, especially in terms of MAP and P@10, even though this difference is less marked than the one reported in the reference paper.

Table 7: TREC 06-07-08 Ad-Hoc test collection. Original and reproduced results for the SNS stemmer and difference between the reproduced stemmer and the original one. Differences greater than 0.0100 are reported in bold.

	Original			Reproduced			Difference		
	NO	RB	SNS	NO	RB	SNS	NO	RB	SNS
MAP	0.2290	0.2599	0.2582	0.2289	0.2596	0.2319	-0.0001	-0.0003	**-0.0263**
RPrec	0.2733	0.3008	0.3001	0.2736	0.3013	0.2722	+0.0003	+0.0005	**-0.0279**
P@10	0.4327	0.4833	0.4727	0.4320	0.4827	0.4267	-0.0007	-0.0006	**-0.0460**

In Table 7 we report the results for the TREC English collection. As in the reference paper, the Porter stemmer performs the best amongst the other approaches and the reproducibility of the baseline strategies (no stemmer and rule-based stemmer) are successfully reproduced. Unfortunately, SNS for TREC presents rather different results from the original paper with a consistent decrease of performances. Despite this drop in performances, SNS still introduced a small improvement in terms of MAP with respect to the no stemmer system even though this is not comparable to the results obtained in the reference paper. These results are quite surprising especially if we consider that for the non-English collections the SNS behavior has been reproduced quite accurately.

GRAS. GRAS was evaluated on four CLEF collections – i.e. the 2006-2007 CLEF Bulgarian and Hungarian collections, the 2007 CLEF Czech collection and the 2005-2006 CLEF French collection – and one TREC collection – i.e. the corpus is the TIPSTER Disk 4&5 minus the congressional record and the federal register and the TREC6, TREC7 and TREC8 topics. The queries are formed of the title and description fields of the topics. The reference paper adopted Terrier, the version of which is not reported.

Table 8: Number of documents and unique words in the considered corpora.

	Original					Reproduced				
	EN	FR	HU	BG	CZ	EN	FR	HU	BG	CZ
Docs	472, 525	177, 452	49, 530	87, 281	81, 735	472, 525	177, 452	49, 530	69, 281	81, 735
Words	522, 381	303, 349	528, 315	320, 673	457, 164	502, 280	325, 292	534, 813	292, 077	457, 149

In the reference paper 87, 281 documents in the Bulgarian corpus are reported, but the number of documents in this corpus is 69, 195 [1], thus we register

a conspicuous difference of 18,086 documents that turn out to produce a lexicon which is 9% smaller than the one of the reference paper. In Table 8 we reported the number of documents and indexed tokens reported by the reference paper and the values we obtained for the same corpora; we can see that the number of documents is the same for all the corpora except the Bulgarian, where the number of unique words differs quite a bit. The greatest differences are recorded for the French where the reference paper reported 7.23% more unique words than we counted and the Bulgarian with 8.92% more unique words. For the English, we have a difference of 3.84%, 1.23% for the Hungarian and less than 0.01% for the Czech language.

GRAS employs a parameter l which is set to be "the average word length for the language concerned"; no further details are given. We chose the values reported in Table 9 by calculating the average word length weighted by their frequency in the considered lexicons. The α and δ parameters have been set up to $\alpha = 4$ and $\delta = 0.8$ as specified in the reference paper.

Table 9: The average word length weighted by their frequency for the considered languages.

	Czech	Bulgarian	English	French	Hungarian
l	6	7	7	7	8

For the reproduction of the experimental results, we focused on the GRAS stemmer and in this case we do not report the results obtained for the no stemmer and the rule-based stemmer cases since they are the same as those reported for the FCB and SNS cases reported above. In the GRAS paper there are no further details about the rule-based stemmers employed thus, even after our reproducibility attempts, we remain uncertain about what stemmer was used for the Hungarian collection.

In Table 10 we report the results we obtained compared to those in the reference paper. By focusing on the MAP values, we can see that we have a sizable difference, between the original and the reproduced stemmer, only for Bulgarian and Hungarian. For Bulgarian, this difference is almost certainly related to the number of documents considered and, consequently, to the size of the lexicon. For Hungarian, the difference between the originally used corpus and the one we employed is quite contained, but as Hungarian is a highly inflected language, it may be more sensitive to the differences related to the lexicon used for training purposes. Another possible reason for these differences may be the selected l parameter, the value of which was not reported in the original paper.

5 Discussion

We considered three statistical stemmers – i.e., FCB, SNS and GRAS – proposed by the same subset of core authors in 2011 and presented in relevant IR venues.

Table 10: Experimental results obtained by reproducing the GRAS stemmer. Differences greater than 0.0100 are reported in bold.

		MAP	R-Prec	P@5	P@10	Rel Ret
BG	Original	0.3260	0.3340	0.4240	0.3550	2110
	Reproduced	0.3410	0.3580	0.4730	0.3720	2043
	Diff	**+0.0150**	**+0.0240**	**+0.0490**	**+0.0170**	−67
CZ	Original	0.3660	0.3600	0.4480	0.3760	689
	Reproduced	0.3630	0.3580	0.4460	0.3720	690
	Diff	−0.0030	−0.0020	−0.0020	−0.0040	+1
EN	Original	0.2700	0.3090	0.5430	0.4790	7873
	Reproduced	0.2749	0.3128	0.5492	0.4859	7904
	Diff	+0.0049	+0.0038	+0.0062	+0.0069	+31
FR	Original	0.3870	0.3980	0.5330	0.4910	4078
	Reproduced	0.3867	0.3886	0.5495	0.4838	4115
	Diff	−0.0003	−0.0094	**+0.0165**	−0.0072	+37
HU	Original	0.3510	0.3600	0.4740	0.4220	1924
	Reproduced	0.3319	0.3467	0.4701	0.4104	1846
	Diff	**-0.0191**	**-0.0133**	−0.0039	**-0.0116**	−78

In some cases, the reproduction of the results reported in the reference papers has been challenging also for the baseline systems where no stemmer or a standard rule-based stemmer were employed.

The considered papers have some pros and cons when it comes to their reproducibility. (i) They employed a standard open-source system as Terrier for the experiments, but they do not report the version used. The use of Terrier limits the number of uncontrolled variables in an experimental setting, but from version to version some key features change and they may impact the reproducibility. (ii) They used standard test collections. This is a good practice of the IR community that enables the comparability of results by guaranteeing that the same corpus, topics and qrels are used; nevertheless, for some collections – e.g., Bulgarian for the GRAS stemmer – the corpus size used in the reference papers does not match the one reported in the CLEF documentation [1]. Moreover, for FCB only the WSJ sub-corpus was employed; this choice requires modification of the official qrels used in the TREC ad-hoc tracks with the possibility of introducing mistakes. (iii) The pseudo-code of the key algorithms is given for two algorithms: SNS and GRAS; this is a good practice because it reduces the ambiguities intrinsic with text descriptions of algorithms.

In general, the considered stemmers have been reproduced quite successfully given that, for at least one test collection per stemmer, we obtained MAP values whose difference with the one reported in the reference papers is less than 0.01. The differences with the reference papers involve the no stemmer and the rule-based stemmer cases and, in most cases, when we failed to reproduce the baseline cases we also found sizable differences with the statistical stemmers – e.g., SNS and GRAS for Bulgarian. In these cases, the difference is possibly due to the

base setting of Terrier rather than to the specific implementation of the statistical stemmer at hand.

Finally, we tested the three stemmers on three common test collections (i.e. Bulgarian, Czech and Hungarian) and we see that GRAS outperforms SNS and FCB for Bulgarian and Czech, whereas SNS outperforms the other two for Hungarian. As expected, for English none of the statistical stemmers outperforms the Porter stemmer, whereas they outperform the rule-based stemmer for the other considered languages, proving their suitability for languages with few linguistic resources.

Acknowledgments
We wish to thank Nicola Ferro who suggested statistical stemmers for a reproducibility study and contributed to discussions along the way.

References

1. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2007: Ad Hoc Track Overview. In: Proc. of the 8th Workshop of the Cross–Language Evaluation Forum (CLEF 2007). pp. 13–32. LNCS 5152, Springer (2008)
2. Dietz, F., Petras, V.: A Component-Level Analysis of an Academic Search Test Collection. - Part I: System and Collection Configurations. LNCS, vol. 10456, pp. 16–28. Springer (2017), https://doi.org/10.1007/978-3-319-65813-1
3. Dolamic, L., Savoy, J.: Indexing and stemming approaches for the Czech language Author links open overlay panel. Information Processing & Management 45(6), 714–720 (2009)
4. Ferro, N., Silvello, G.: CLEF 15th Birthday: What Can We Learn From Ad Hoc Retrieval? In: Proc. of the 5th International Conference of the CLEF Initiative (CLEF 2014). pp. 31–43. LNCS 8685, Springer (2014)
5. Krovetz, R.: Viewing Morphology as an Inference Process. In: Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993). pp. 191–202. ACM Press (1993)
6. Lovins, J.B.: Development of a Stemming Algorithm. Mechanical Translation and Computational Linguistics 11(1/2), 22–31 (March/June 1968)
7. Macdonald, C., McCreadie, R., Santos, R.L.T., Ounis, I.: From Puppy to Maturity: Experiences in Developing Terrier. Proc. of OSIR at SIGIR pp. 60–63 (2012)
8. Paik, J.H., Mitra, M., Parui, S.K., Järvelin, K.: GRAS: An effective and efficient stemming algorithm for information retrieval. ACM Trans. Inf. Syst. 29(4), 19 (2011)
9. Paik, J.H., Pal, D., Parui, S.K.: A Novel Corpus-based Stemming Algorithm Using Co-occurrence Statistics. In: Proc. 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011). pp. 863–872. ACM Press (2011)
10. Paik, J.H., Parui, S.K.: A Fast Corpus-Based Stemmer. ACM Trans. Asian Lang. Inf. Process. 10(2), 1–16 (2011)
11. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (July 1980)
12. Savoy, J.: Searching strategies for the Hungarian language. Information Processing & Management 44(1), 310–324 (2008)
13. Singh, J., Gupta, V.: Text Stemming: Approaches, Applications, and Challenges. ACM Computing Surveys (CSUR) 49(3), 45:1–45:46 (2016)