A Systematic Comparison of 3 Phrase Sampling Methods for Text Entry Experiments in 10 Languages

Germán Sanchis-Trilles
PRHLT Research Center
Universitat Politècnica de València
gsanchis@prhlt.upv.es

Luis A. Leiva
PRHLT Research Center
Universitat Politècnica de València
luileito@prhlt.upv.es

Work partially supported by EU FP7 program under grant agreements 287576 (CASMACAT) and 600707 (tranScriptorium).

MobileHCI'14, Sep 23–26 2014, Toronto, ON, Canada
ACM 978-1-4503-3004-6/14/09.
http://dx.doi.org/10.1145/2628363.2634229

Abstract
Today's reference datasets for conducting text entry experiments are only available in English, which may lead to misleading results when testing text entry methods with non-native English speakers. We compared 3 automated phrase sampling methods available in the literature: RANDOM, NGRAM, and MEMREP. It was found that MEMREP performs best according to a statistical analysis and qualitative observations. This resulted in a collection of 30 datasets across 10 major languages, and we wish to share them with the community via this paper.

Author Keywords
Text Entry; Sampling; Memorability; Representativeness

ACM Classification Keywords
H.5.2 [**Information interfaces and presentation**]: *Input devices and strategies, Evaluation/Methodology*

Introduction
Text entry is perhaps the task that users perform the most on a mobile device. In fact, there is a wealth of research on text entry methods through development and evaluation. Among these two fronts, in this paper we are interested in the latter, in order to better understand and improve them both.

The most popular methodology for evaluating text entry methods, whether on a mobile device or not, is to ask participants to transcribe a predefined set of phrases (short sentences). This methodology is interesting for a number of reasons [8]. First, a transcription task eliminates noise by reducing the variability that might occur otherwise if users were allowed to enter free text. Second, if the phrases used as stimuli are always the same for all participants, this facilitates the comparison of different text entry techniques. Third, a transcription task does not require participants to create their own text, which removes additional cognitive processing time. Last, and most important, this allows results to be reproducible.

It may seem more natural to have users enter free text in order to increase the external validity of the experiment. However, it is critical to make the text entry method the only independent variable in the experiment, i.e., increasing its internal validity. Then, the question of how to present the phrases to the user still prevails. In general, copy-tasks should prefer memorable stimuli [3, 4, 7]. This is important to avoid participants consulting often the reference text, which shifts attention away from the text entry method.

In this paper, we provide 30 datasets (of 2,000 phrases each) across 10 major languages. These datasets are publicly available and are intended to be a reference source to conduct text entry experiments in languages other than English, though we included English phrase sets in order to allow others to compare their overall performance. We performed a systematic comparison of all datasets according to memorability and representativeness, two desired properties for conducting mobile text entry experiments. We conclude that MemRep is the most adequate phrase sampling method.

Related Work

For the past decade, text entry researchers have predominately used the MacKenzie and Soukoreff dataset [4], which contains 500 phrases that were manually selected according to three criteria: moderate in length, easy to remember, and representative of general English. More recently, Vertanen and Kristensson [7] released a phrase set based on genuine mobile emails. Mobile text entry method evaluations should use this dataset because it consists of memorable and representative phrases taken from actual mobile email messages. Moreover, this dataset has been carefully reviewed so that phrases are well-formed and different subsets are suitable for use in a variety of text entry evaluations.

Unfortunately, these popular datasets are only available in English. In contrast, today text is entered into mobile devices in many different languages, where text entry methods might perform very differently; c.f., English vs. Polish vs. Russian. This may lead to misleading results when testing text entry methods with non-native English speakers, as it has been shown that task performance is highly influenced by language proficiency [2, 3].

Paek and Hsu [5] devised a procedure for creating representative phrase sets by randomly sampling sets of n-grams (sequences of n words) and choosing the set with less entropy with regard to the original dataset. Being data-driven, this procedure allows to generate phrase sets in potentially any language. Finally, Leiva and Sanchis-Trilles [3] presented an automated phrase sampling method that takes into account memorability and representativeness. This method has been shown to work well for English and Spanish, though it has not been explored in other languages.

Language	Corpus	Size
Dutch	Wikipedia	1.2
	OpenSubtitles	3.7
English	Wikipedia	6.9
	OpenSubtitles	14.0
French	Wikipedia	2.5
	OpenSubtitles	3.9
German	Wikipedia	3.6
	OpenSubtitles	0.8
Italian	Wikipedia	1.8
	OpenSubtitles	2.0
Polish	Wikipedia	1.1
	OpenSubtitles	7.3
Portuguese	Wikipedia	1.0
	OpenSubtitles	4.1
Russian	Wikipedia	2.7
	OpenSubtitles	1.5
Spanish	Wikipedia	2.2
	OpenSubtitles	6.8
Swedish	Wikipedia	0.8
	OpenSubtitles	1.1

Table 1: Size of input datasets in GB. Note that the corpora are fairly large, and processing them is very costly, e.g., processing the English data required around 0.5 TB RAM using MapReduce. Hence, we hope that having these sets readily available will foster cross-lingual text entry research.

Phrase Sampling Methods

Our datasets were compiled using 3 different sampling techniques for text entry experiments: RANDOM, NGRAM, and MEMREP. To our knowledge, these techniques are the only ones within the literature that are completely automated, and can be used to build datasets that are comparable across languages and domains.

The RANDOM method relies on selecting phrases at random from the input dataset. This is the most simple form of phrase selection. If the sample obtained is sufficiently large, then representativeness is ensured. However, this method does not take into account phrase memorability.

The NGRAM method [5] focuses on a more precise measure of representativeness, as computed according to perplexity (cross-entropy). One limitation of this technique is that it does not sample complete phrases, but n-grams of fixed length. Moreover, it does not take into account the memorability of the sampled n-grams.

The MEMREP method [3] computes memorability based on a number of phrase features selected empirically. Representativeness is taken into account by weighting phrases with a probability density function over the phrase features. This method relies on having two different input datasets: one which is considered large enough to be descriptive of general language, and one from which phrases will be selected.

Datasets

Ideally, phrases for mobile text entry experiments should be sampled from publicly available email or SMS datasets. However, mainly for privacy reasons, the public release of these kind of data is rare [7]. In fact, we are not aware of any dataset composed of actual mobile messages in languages different from English. Therefore, we considered the OpenSubtitles 2011 dataset (see [6] for details) for ten major languages (Tables 1 and 2). Previous work [1] has shown that this kind of texts are likely to be memorable, a desired property for conducting text entry experiments. Phrase repetitions were removed prior to sampling.

Language	Sentences	Word count	Voc. Size	Singletons
Dutch	127.8 M	737 M	1,018.8 k	116.9 k
English	265.1 M	1,575 M	812.3 k	244.1 k
French	129.8 M	781 M	694.2 k	111.2 k
German	25.1 M	137 M	611.4 k	106.7 k
Italian	65.7 M	376 M	610.1 k	55.9 k
Polish	183.9 M	843 M	1,113.5 k	108.0 k
Portuguese	141.0 M	764 M	684.9 k	73.0 k
Russian	28.9 M	144 M	962.3 k	227.0 k
Spanish	212.1 M	1,219 M	1,153.2 k	184.0 k
Swedish	38.1 M	200 M	660.2 k	98.7 k

Table 2: Overview of the Subtitles corpus for all languages.

Since the MEMREP technique requires an additional dataset as input to model the knowledge of each language, we downloaded the latest Wikipedia articles (May 2014) for those languages having at least 1 million articles (Tables 1 and 3). This large amount of data can be considered representative enough of a language; c.f. vocabulary sizes in Tables 2 and 3. However, because MEMREP relies on some character-based features, we excluded those languages where words are ideograms, e.g., Chinese. This being done, the eventually selected languages were: Dutch, English, French, German, Italian, Polish, Portuguese, Russian, Spanish, and Swedish. We believe this language selection will comprise a big part of the languages that might be subject of study in mobile text entry experiments. The datasets are available at `http://personales.upv.es/luileito/memrep/`

Language	Sentences	Word count	Voc. Size	Singletons
Dutch	7.7 M	192 M	2.6 M	1.1 M
English	63.2 M	1,232 M	5.8 M	3.0 M
French	13.6 M	428 M	2.7 M	1.3 M
German	16.0 M	534 M	7.3 M	3.9 M
Italian	9.2 M	300 M	2.3 M	1.1 M
Polish	6.1 M	156 M	2.5 M	1.1 M
Portuguese	5.2 M	165 M	1.7 M	0.8 M
Russian	6.9 M	20 M	4.1 M	2.1 M
Spanish	10.1 M	374 M	2.6 M	1.3 M
Swedish	5.5 M	114 M	2.4 M	1.1 M

Table 3: Overview of the Wikipedia corpus for all languages.

Analysis

Table 4 provides an overview of the phrases sampled by each technique and for all languages. Three interesting observations can be appreciated. First, even though the average number of words per phrase are similar, RANDOM produces more variability (min. *SD*=3.7 in Swedish and max. *SD*=5.3 in Italian). This reveals that this method may pick very long sentences, and therefore we discourage its use for conducting mobile text entry experiments.

Second, MEMREP tends to produce short words of similar length (min. *SD*=1.5 in Swedish and max. *SD*=2.0 in Italian), whereas the other techniques do not; e.g., min. *SD*=2.7 for RANDOM and min. *SD*=2.8 for NGRAM. This suggests that both RANDOM and NGRAM tend to produce very long (but also very short) words, which overall will be harder to memorize.

Third, the 3 sampling methods produce vocabularies of similar length, around 2k and 3k words per language, although the MEMREP vocabularies are consistently 1k smaller and, more importantly, it is the only method where *all* phrases are comprised of frequent words.

Further analysis is shown in Figure 1, where we plot the density functions of memorability and representativeness, according to previous work [3]. It can be observed that both RANDOM and NGRAM present very disperse curves, which suggests that highly memorable and representative phrases are mixed along with others which are neither representative nor memorable. This dispersion should not be underestimated, since it might introduce undesired variability into a text entry experiment and thus might compromise its internal validity. On the contrary, MEMREP presents a very sharp peak, both in representativeness and in memorability. Hence, all of the selected phrases, although different, are statistically similar in their properties.

We conclude that MEMREP is the most adequate sampling method to generate phrase sets for mobile text entry experiments. Examples of the selected phrases are shown in Figure 2.

Language	Words/Phrase			Letters/Word			OOV Ratio		
	R	N	M	R	N	M	R	N	M
Dutch	5.7	4.8	5.0	4.3	4.1	4.12	6.0 %	0.8 %	0 %
English	6.0	4.8	4.0	3.8	3.8	3.55	1.4 %	0.2 %	0 %
French	5.9	4.8	5.0	4.1	4.0	3.81	3.4 %	0.7 %	0 %
German	5.5	4.7	4.0	4.8	4.7	4.39	5.7 %	1.0 %	0 %
Italian	5.8	4.8	4.0	4.6	4.5	4.15	6.3 %	1.5 %	0 %
Polish	4.6	4.7	4.0	5.0	4.9	4.18	11.9 %	2.7 %	0 %
Portuguese	5.4	4.7	4.0	4.3	4.3	4.01	6.2 %	1.1 %	0 %
Russian	5.0	4.7	4.0	4.8	4.7	4.04	11.1 %	1.9 %	0 %
Spanish	5.8	4.8	4.0	4.4	4.3	4.43	5.8 %	1.1 %	0 %
Swedish	5.4	4.7	4.0	4.2	4.2	3.85	6.7 %	1.6 %	0 %

Table 4: Overview of the 2,000 phrases selected by RANDOM (R), NGRAM (N) and MEMREP (M). OOV stands for Out-Of-Vocabulary, i.e., unique words that did not appear in the Wikipedia corpus.

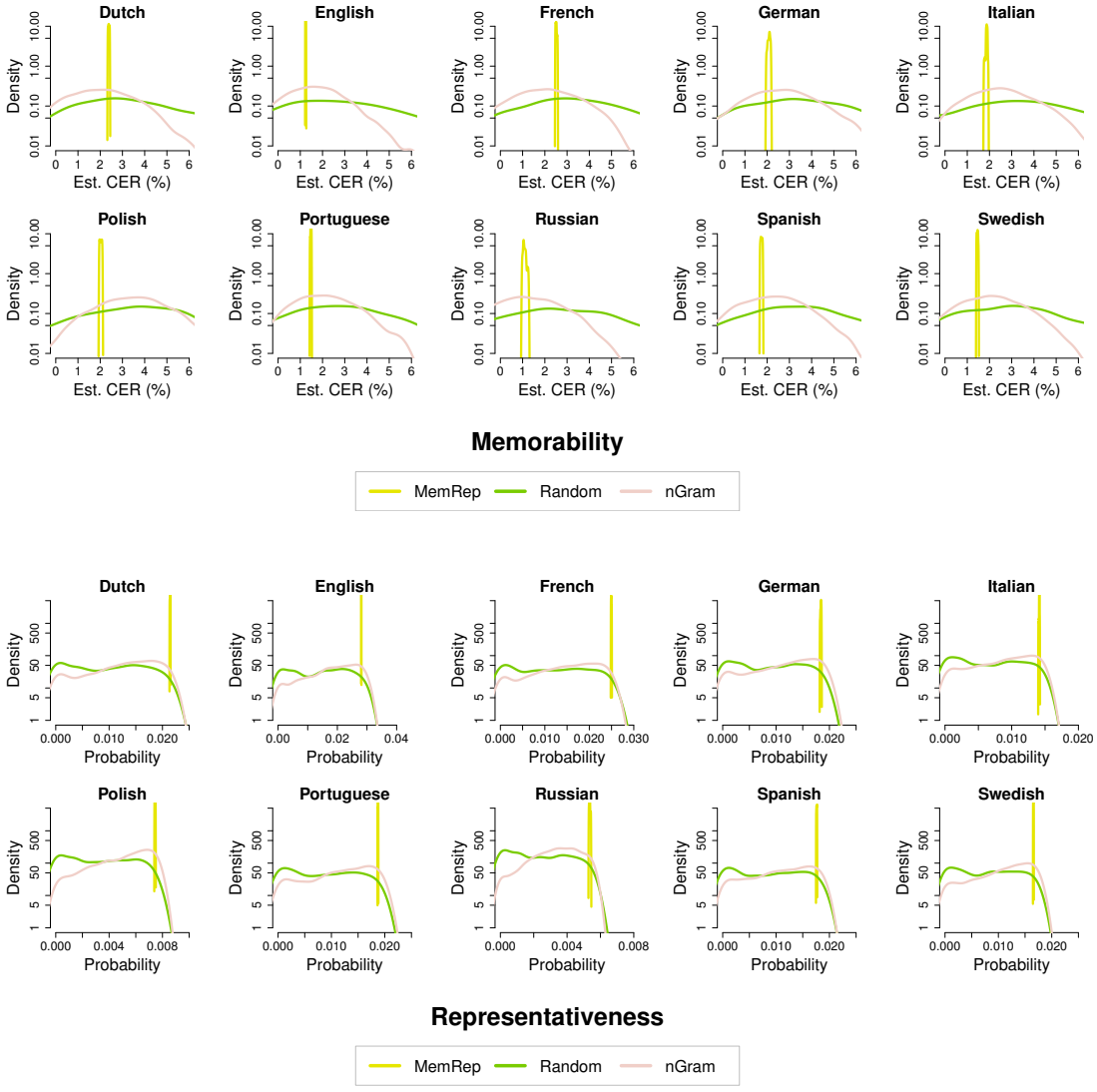

Figure 1: Dataset distributions overview. MemRep phrases are consistently concentrated around more desirable values of memorability (low estimated character error rate, top row) and representativeness (high probability, bottom), according to [3].

Conclusion and Future Work

We have collected new datasets for conducting mobile text entry experiments in 10 major languages. All datasets are statistically comparable, however we recommend the MemRep datasets because they are highly memorable and representative of each language. A limitation of our work is that the datasets have not undergone full human inspection. However, the statistical analysis conducted suggests that their behavior will be similar across all languages, and MemRep has already been assessed with English and Spanish speakers [3]. Nevertheless, in future work we will test all datasets with their native speakers.

References

[1] Danescu-Niculescu-Mizil, C., Cheng, J., Kleinberg, J., and Lee, L. You had me at hello: How phrasing affects memorability. In *Proc. ACL* (2012), 892–901.

[2] Isokoski, P., and Linden, T. Effect of foreign language on text transcription performance: Finns writing English. In *Proc. NordiCHI* (2004), 109–112.

[3] Leiva, L. A., and Sanchis-Trilles, G. Representatively memorable: Sampling the right phrase set to get the text entry experiment right. In *Proc. CHI* (2014), 1709–1712.

[4] MacKenzie, I. S., and Soukoreff, R. W. Phrase sets for evaluating text entry techniques. In *Proc. CHI EA* (2003).

[5] Paek, T., and Hsu, B.-J. P. Sampling representative phrase sets for text entry experiments: a procedure and public resource. In *Proc. CHI* (2011), 2477–2480.

[6] Tiedemann, J. Parallel data, tools and interfaces in OPUS. In *LREC* (2012), 2214–2218.

[7] Vertanen, K., and Kristensson, P. O. A versatile dataset for text entry evaluations based on genuine mobile emails. In *Proc. MobileHCI* (2011), 295–298.

[8] Vertanen, K., and Kristensson, P. O. Complementing text entry evaluations with a composition task. *ACM TOCHI 21*, 2 (2014), 8:1–8:33.

Language	Method	Sample phrases
Dutch	RANDOM	oké niet huilen
	NGRAM	getuiges in een belangrijke zaak als ik jou niet afmaak
	MEMREP	grapje het is gewoon bier zo heette mijn moeder ook
English	RANDOM	and i ain't so afraid of losing something [8+] pipe it
	NGRAM	we intend to escape he can strip naked and
	MEMREP	the future seemed bright at the lemon slice
French	RANDOM	non t'es sur la liste à la réputation peu reluisante
	NGRAM	oublié de l'interroger sur depuis que vous avez conçu
	MEMREP	nous restons avec le peuple parfois je blague avec elle
German	RANDOM	jemanden mit einem 100 000 wagen [7+] 651 01 07 33 652 gt 01 07 35 210 ja i m [2+]
	NGRAM	marie die mir den kopf waren vor 100 jahren abhängig
	MEMREP	hab ich auch gehört mein antrieb ist weg
Italian	RANDOM	robin che lavora con gisborne potrebbe [2+] scusi ho sbagliato strada
	NGRAM	e tutto cosi confortevole e sta facendo un inchiesta per
	MEMREP	e' tanto grave questa sabbia si sposta

Language	Method	Sample phrases
Polish	RANDOM	ohydny dobry jest
	NGRAM	jakby pisał z myślą o że może mieć grypę
	MEMREP	teraz nie ma niczego nie są za pomocni
Portuguese	RANDOM	moradores são mais que bem vindos bebe do sapato
	NGRAM	que faça uma argumentação convencida de que é o
	MEMREP	não não matou ninguém bem pessoal acho que
Russian	RANDOM	ложись дэнни почему все так трясутся когда заходит речь [4+]
	NGRAM	подобное с другим человеком жизни и ты не хочешь
	MEMREP	не пускай их сюда он придет за мной
Spanish	RANDOM	bien lo admito es un cocodrilo centrai regreso con ustedes despues de caminar [2+]
	NGRAM	no bromearía con una cosa tu cara que ibas a
	MEMREP	así lo hacemos todos vuela como un pájaro
Swedish	RANDOM	det är miss och vivian bättre än på både violet och grace
	NGRAM	jag drömmar framträda och drömmar nej jag menade inte
	MEMREP	det finns alltså nån vi gjorde dock inget

Figure 2: Examples of the selected phrases, extracted at random from each dataset. [N+] means that the phrase contains N more words, which have been removed here for space reasons. The MEMREP datasets are available in three forms: full phrases, punctuation symbols removed, and tokenized + lowercased + no punctuation (as in this figure). The other datasets are provided for replicability and therefore are only available in tokenized + lowercased + no punctuation form.