Int. J. Human-Computer Studies 113 (2018) 15–31

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

# Multilingual phrase sampling for text entry evaluations

Marc Franco-Salvador[a], Luis A. Leiva[b,*]

[a] Symanto Research, Pretzfelder Strasse 15, Nuremberg DE-90425, Germany
[b] Sciling, SL, Camí a la Mar, 75, Valencia 46120, Spain

ARTICLE INFO

*Keywords:*
Phrases
Sentences
Sampling
Multilingualism
Memorability
Representativeness
Semantics

ABSTRACT

Text entry evaluations are typically conducted with English-only phrase sets. This calls into question the validity of the results when conducting evaluations with non-native English speakers. Automated phrase sampling methods alleviate this problem, however they are difficult to use in practice and do not take into account language semantics, which is an important attribute to optimize. To achieve this goal, we present KAPS, a phrase sampling method that uses the BabelNet multilingual semantic network as a common knowledge resource, aimed at both *standardizing* and *simplifying* the sampling procedure to a great extent. We analyze our method from several perspectives, namely the effect of sampled phrases on user's foreign language proficiency, phrase set memorability and representativeness, and semantic coverage. We also conduct a large-scale evaluation involving native speakers of 10 different languages. Overall, we show that our method is an important step toward and provides unprecedented insight into multilingual text entry evaluations.



## 1. Introduction

Text entry as a research discipline is increasingly attracting the interest of many researchers. By way of example, as of January 2017 the query (“text entry” “text input”) returns 647 results in the ACM digital library, and we can see that the number of publications has doubled each lustrum in the last 15 years. These figures put forward the fact that text entry is a rapidly growing community.

Some authors argue that text entry research has seen a revival in recent years due to the advent of mobile devices. Indeed, academic and industry researchers have been working on text entry since the emergence of handheld technologies (Dunlop and Masters, 2009; Wobbrock and Myers, 2006). Eventually, as any other interaction technique, text entry methods need to be evaluated. However, it is well-known that the outcome of text entry experiments is affected by the text users enter (Mackenzie and Felzer, 2010).

### 1.1. Text entry evaluations

Typically, in text entry experiments participants are prompted with phrases (short sentences) that must be entered as quickly and accurately as possible. Phrases can be full sentences or sentence fragments such as greetings, idioms, or quotations. An alternative option to *transcription* (i.e., copying pre-selected text) is *composition* (i.e., generating new text). This is considered more ecologically valid when evaluating text entry techniques (Zhai et al., 2005), since it mimics typical device usage. However text composition tasks are more difficult to control and measure. For example, since there is no reference text available, it is not possible to compute well-established measures of input error such as the character error rate. As a workaround, Vertanen and Kristensson (2014) described how human judging can provide a surrogate error rate measure that ensures participants are making good faith efforts to be accurate.

Overall, although it may seem more natural to have users enter free text and increase thus the external validity of the experiment (i.e., the extent to which the observed effect is generalizable), it is critical to make the text entry method the only independent variable in the experiment, and increase thus its internal validity (i.e., the extent to which the observed effect is due to the test conditions). Indeed, if users were asked to “type anything as fast as possible” they would introduce rather biased (maybe nonsensical) text. Hence, text entry researchers typically use pre-selected phrases, measuring the dependent variables (e.g., input speed or error rates) in a transcription task. This eliminates noise and facilitates the comparison of text input techniques across studies (Leiva and Sanchis-Trilles, 2014; Vertanen and Kristensson, 2011b).

In general, transcription tasks should prefer memorable stimuli (Leiva and Sanchis-Trilles, 2014; MacKenzie and Soukoreff, 2002; Vertanen and Kristensson, 2011b). This reduces participants' tendency to shift attention between the stimulus phrase and the text entry method. To ensure memorable stimuli, researchers often resort to using manually curated English-only phrase sets, which are typically small according to

* Corresponding author.
*E-mail addresses:* marc.franco@symanto.net (M. Franco-Salvador), name@sciling.com (L.A. Leiva).

https://doi.org/10.1016/j.ijhcs.2018.01.006
Received 3 February 2017; Received in revised form 19 January 2018; Accepted 22 January 2018
Available online 31 January 2018
1071-5819/

modern standards, or rely on sampling procedures that do not guarantee the internal validity of the experiment. In contrast, today text is entered into many different devices in many different languages, where text entry methods might perform very differently. This fact evidences the necessity of an adequate phrase sampling method, aimed at exploiting the huge amount of text corpora available in different languages.

Currently, most text entry experiments in HCI involving English users use either the phrase set released by MacKenzie and Soukoreff (2003) (MacKenzie dataset for short) or the EnronMobile dataset (Vertanen and Kristensson, 2011b). Both phrase sets have been proved to be adequate for conducting experiments with English participants. Furthermore, using these sets make it easier to reproduce different studies conducted by other researchers. The problem, however, is how to conduct text entry experiments with non-English users or in very specialized fields (e.g. a text entry method for a medical device, where technical vocabulary is commonplace), or simply when the participants speak different languages.g It is here where automated sampling methods like Ngram (Paek and Hsu, 2011) or MemRep (Leiva and Sanchis-Trilles, 2014) are valuable. On the one hand, the Ngrammethod approaches phrase sampling as a single-objective function, which only considered the representativeness of the phrases as the measure to optimize (to be described later). On the other hand, the MemRep sampling method approaches phrase sampling as a dual-objective function, incorporating both representativeness and memorability as the measures to optimize (also to be described later).

To the best of our knowledge, to date MemRep is the only automated method that provides adequate phrases for conducting text entry experiments in languages different from English. However, MemRep is cumbersome to use in practice, since it requires an additional large dataset to learn a language model, in addition to the input dataset from which phrases will be sampled. Moreover, MemRep does not take into account phrase semantics, which may result in confusing phrases like '`Send e.m.s.`' or '`100 is proposed for this category`'. We elaborate this discussion in the next sections.

### 1.2. Contributions

In this article, we present Kaps (acronym for Knowledge-Aided Phrase Sampling), an automated method for sampling phrase sets that uses knowledge graphs to select phrases for text entry seeking a balance among memorability, representativeness, and semantics. Our method is based on a multiple regression model over language-independent features, so that it can generalize to other languages. Kaps uses the BabelNet multilingual semantic network as a common resource, for *standardization* purposes, which also *simplifies* the evaluation procedure to a great extent. For example, contrary to MemRep , Kaps does not need a statistical analysis of an additional (large) corpus, just the dataset from which phrases will be drawn. An interesting property of our method is that, being data-driven, the sampled phrases are prototypical of the language or domain of interest.

We analyze Kaps from several perspectives, namely the effect of sampled phrases on user's foreign language proficiency, phrase set memorability and representativeness, and semantic coverage. We also conduct a large-scale evaluation involving 200 native speakers in 10 different languages. Overall, we show that our method is an important step toward and provides unprecedented insight into multilingual text entry evaluations. Finally, we make our software and data (ready-made phrase sets) publicly available so that others can build upon our work.

## 2. Related work

The choice of phrase set has been extensively discussed in the text entry literature. In the past, researchers used ad-hoc text sources for their experiments, such as sentences drawn from a Western novel (Karat et al., 1999), quotations from Unix's `fortune` program (Isokoski and Raisamo, 2000), news snippets (Zhai et al., 2002), street addresses (González et al., 2007), or passages from Sherlock Holmes (Tanaka-Ishii et al., 2003) and Alice in Wonderland (Vasiljevas et al., 2015). Using ad-hoc, proprietary text sources is often considered a bad practice because text entry studies could not be accurately reproduced. To help the situation, (MacKenzie and Soukoreff, 2003) released a phrase set consisting of 500 phrases that was also designed to contain easy to remember text. However, the MacKenzie phrase set mainly consists of short English idioms and clichés, and the memorability of the phrase set was never verified. Vertanen and Kristensson (2011b) processed the Enron email dataset and released the EnronMobile phrase set, including empirical data regarding sentence memorability.

**Table 1**
Survey of recent research on text entry involving user studies, according to the ACM digital library.

| Language | No. Studies | Dataset Used | | |
|---|---|---|---|---|
| | | MacKenzie | EnronMobile | Custom |
| English | 91 | 55 | 10 | 26 |
| French | 15 | 7[a] | | 8 |
| Finnish | 14 | 11[b] | | 3[c] |
| Portuguese | 12 | 2[d] | 1[d] | 9 |
| German | 12 | 5 | 1 | 6 |
| Korean | 8 | 5[e] | | 3 |
| Hindi | 7 | 2 | | 5 |
| Chinese | 6 | 1 | | 5 |
| Spanish | 5 | 2 | | 3[f] |
| Japanese | 4 | 2 | | 2 |
| Dutch | 3 | 1 | | 2 |
| Italian | 2 | | | 2 |
| Africans | 2 | 1 | | 1 |
| Norwegian | 1 | 1 | | |
| Bulgarian | 1 | | | 1 |
| Arabic | 1 | 1 | | |
| Myanmar | 1 | | | 1 |
| Bengali | 1 | | | 1 |
| Total | 185 | 96 | 12 | 78 |

[a] 2 were translated to French.
[b] 7 were translated to Finnish.
[c] 1 was left in English.
[d] 2 were translated to Portuguese.
[e] 2 were translated to Korean.
[f] 1 study used also the original MacKenzie dataset.

Both MacKenzie and EnronMobile are today the most popular phrase sets used in text entry experiments. Kristensson and Vertanen (2012) compared both phrase sets and found not much difference between them, although the actual differences are *conceptually* rather large. For example, EnronMobile is better suited to evaluating mobile text entry methods, as it contains genuine mobile emails. Other researchers have developed alternative phrase sets for specialist applications. For example, Kano et al. (2006) curated a phrase set for specific use with children and Vertanen and Kristensson (2011a) created a phrase set for Augmentative and Alternative Communication (AAC) by using messages suggested by AAC specialists.

### 2.1. Multilingual text entry

It has been argued that the choice of phrase set might not matter much as long as it is memorable and somewhat representative of the text users write (Vertanen and Kristensson, 2011b). However, current standard datasets for text entry are only available in English. Meanwhile, it is clear and obvious that "text entry" does not imply "English text entry" (MacKenzie and Soukoreff, 2002). Many text entry researchers are conducting user studies in many languages different from English; see Table 1.

Text entry is therefore multilingual, and not being aware of this fact may be problematic, evaluation-wise. In India, for example, 25% of the urban and 64% of the rural mobile phone users speak in Vernacular languages and not in default English (Ghosh and Joshi, 2014). As can be noticed in Table 1, near half of the recent user studies on text entry (42%) have used a custom phrase set. Furthermore, if we look at the non-English studies, this number increases to a significant 55%.

As a matter of fact, transcription tasks have been found to be language-sensitive. For example, Költringer et al. (2007) stated that "participants had good English reading and writing skills, but they were not native English speakers". Isokoski and Linden (2004) noticed that participants' language proficiency affects text entry rates. Leiva and Sanchis-Trilles (2014) provided empirical evidence that the choice of the phrase set affects both memorability and error rates when phrases are not shown in the native language of the participants. These observations suggest that, when conducting text entry experiments with non-native English speakers or in very specialized domains, we either use a standard phrase set and accept that there will be differences in performance across studies, or we have to develop language- or domain-specific phrase sets. To solve this problem, automated phrase sampling methods become necessary.

### 2.2. Automated approaches to phrase sampling

The simplest approach to automatically create phrase sets is by randomly sampling a large text corpus $\mathcal{U}$ and keeping $N$ phrases ($N \ll |\mathcal{U}|$). The problem with this approach is that the outcome is unpredictable and thus it might be irreproducible across studies. The natural language processing and information retrieval communities have devised many methods to select "representative" sentences from text sources. For example, methods based on readability (Mostafazadeh et al., 2011), semantic coherence (Ferret, 1998; Misra et al., 2008), connectivity between sentences (Yoshimi et al., 1998), topic clustering (Radev et al., 2000; Zajic et al., 2006) and topic change detection (Prince and Labadié, 2007), or viewpoint preservation (Zhu et al., 2013). The drawback of these methods is that they are tailored to document summarization and categorization, so they basically aim for removing redundancy or maximizing coverage within a text document or collection thereof. Therefore, it is unclear if they could be directly used to sample phrases that are adequate to conduct text entry evaluations. For example, they do not consider phrase memorability, which is a fundamental property and so it must be ensured.

In the text entry community we can find two competitive approaches to automatically create phrase sets. Paek and Hsu (2011) proposed an entropy-based sampling method to generate representative $n$-grams (fixed length phrases) from a sufficiently large and general text corpus $\mathcal{U}$. They proceeded by creating a random number of phrase subsets of size $N$ from $\mathcal{U}$ and keeping the subset with lower distance (Kullback-Leibler divergence) with regard to $\mathcal{U}$. This method was proved insufficient by Leiva and Sanchis-Trilles (2014), who proposed MemRep , focused on sampling representative *and* memorable phrases. They proceeded by ranking an input phrase set $S \nsubseteq \mathcal{U}$ according to a memorability model that was compensated with the representativeness of $\mathcal{U}$, and keeping the $N$ best input phrases in the rank. MemRep was evaluated with end users and outperformed the previous automated methods (NgramAnd random sampling).

Many text entry methods require constant visual attention (e.g. eye typing) or involve multitasking (e.g. dialing a contact while walking). For experiments trying to emulate similar situations, memorability is critical since it can be difficult for participants to consult often (or even hear) the reference text. Memorability is also desirable to unburden the participants and let them exclusively focus on the text entry method. Previous sampling methods aimed to select representative phrases, but memorability was largely ignored. To our knowledge, MemRep is the only automated method that ensures memorability by design. However, MemRep is difficult to use in practice because it requires 2 different datasets as input: a large dataset from which phrases will be drawn and an even larger dataset to model the language in which phrases are written. This additional dataset is required to compute a series of lexical features, namely phrase probability and out-of-vocabulary (OOV) ratio, so the choice of such additional dataset is critical for MemRep . For example, it cannot be a superset of the dataset used to sample phrases from, otherwise no word would be considered infrequent and the OOV ratio would be 0% for all phrases, leading to wrong model estimates.

Moreover, MemRep does not take into account phrase semantics, which is an important attribute that contributes to ensuring memorability (Rubin, 1977; Thorn and Page, 2009). Although defining what makes an phrase memorable is subtle. Memorability is not simply a question of "memorizability", as it relates to psychological studies of both short and long-term recall (Bates et al., 1980; Danescu-Niculescu-Mizil et al., 2012). Memorability is also subjective, and largely dependent on users' language proficiency. Keller (2004) showed that processing effort is correlated with word length and word frequency. In general, shorter and frequent words take less time to read and therefore are easily understandable (Just and Carpenter, 1980). As discussed later, these observations are key to our work.

In this article, we show that it is possible to use adequate phrases in different languages, ensuring that such phrases will behave similarly across multilingual text entry evaluations. This is important because some languages use more characters per word than other languages[1] and thus a text entry evaluation would report an artificially lower text entry throughput. We should emphasize that our goal is not to replace the traditional English phrase sets but to provide researchers with a reliable automated sampling method that can work across languages and/or domains.

## 3. Method

We are interested in a sampling method to select, from a large text corpus, those phrases that are good candidates for conducting text entry experiments. Such a method should select phrases that take into account the following properties:

1. **memorability**, for ensuring the internal validity of the experiment;
2. **representativeness** of the task, domain, or language of interest, for ensuring external validity;
3. **complexity**, which agglutinates the following properties:
   (a) **syntactics**, since ill-formed or fragmented phrases may be confusing to participants;
   (b) **semantics**, because phrases with no meaning may as well be confusing.

Notice that some of these desired properties might resonate with each other. For example, phrases that are very short will be very memorable, but will rarely be representative either of general language or a particular text entry task or domain. For example, a dataset containing only phrases like '`I am`', '`hi`', or '`and so on`' would not be very appropriate for conducting a text entry experiment. Also consider that semantics dominates over syntactics, since it is possible to infer syntactic cues from semantic cues but not the other way around (Naseem and Barzilay, 2011). Furthermore, phrases can be syntactically correct but semantically incorrect; consider e.g. the phrase '`colorless green ideas sleep furiously`' coined by Chomsky (1955) to offer proof that "grammar is not a valid structure underlying language, rather than words are symbols with associated properties that will not function if they are not properly used". Therefore, we should devise a method that takes into account the above-mentioned properties altogether, seeking a reasonable balance among them.

With the desired properties of a phrase sampling method in mind (high memorability, high representativeness, and low complexity), we

[1] http://www.ravi.io/language-word-lengths

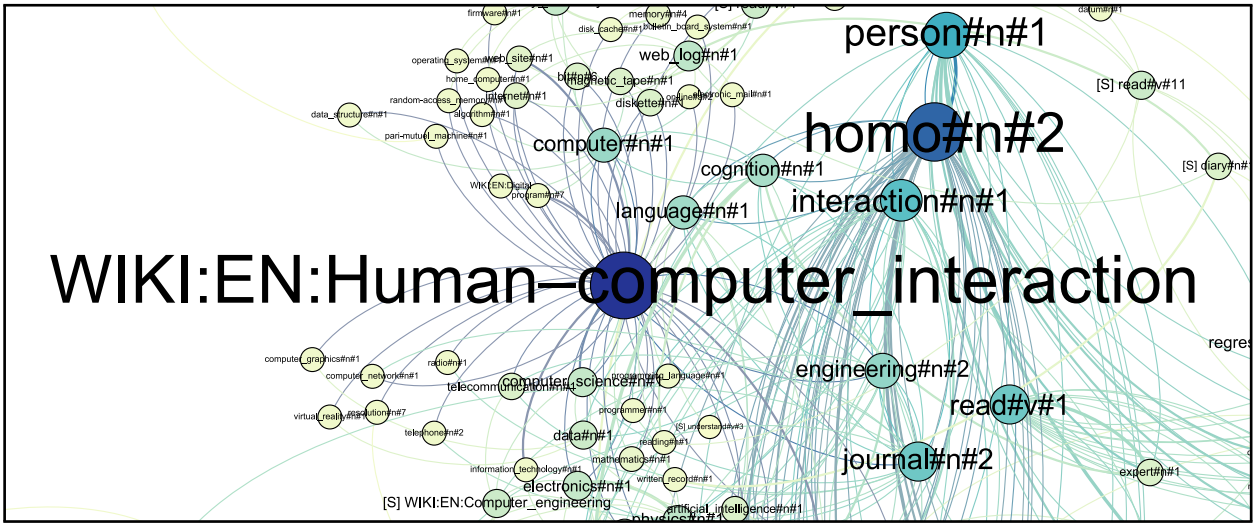


**Fig. 1.** Knowledge graph (fragment) built from the phrase '`I'm reading a journal about human-computer interaction`'. Graph nodes represent concepts. Larger nodes indicate concepts with higher connectivity.

find it important to devise an abstract representation of words. After all, the same concept often can be written in different ways and it happens in many different languages. Moreover, a common concept representation becomes necessary to ensure that our phrase sampling method will be consistent across languages. To achieve this, we will use knowledge graphs derived from a large knowledge base (as described next). In fact, the use of a knowledge base has been used in previous works on text entry. For example, Stocky et al. (2004) showed comparable performance of knowledge graphs to other statistical methods for word prediction. Furthermore, incorporating syntactic information has shown improvements in terms of error rate reduction that reach 30% (Ganslandt et al., 2009). We have named our method KAPS (acronym for Knowledge-Aided Phrase Sampling) after the knowledge graphs notion, described below.

### 3.1. Knowledge graphs: a unified representation of phrase complexity

A knowledge graph is a weighted directed graph that expands and relates the concepts belonging to a text (James, 1992). Notice that words may be polysemous, i.e., with multiple senses or meanings. Similarly, a sense may be represented with different words, specially across languages. A "concept" is the language-independent and abstract representation of one sense, which in this work is represented by a *synset*; see Section 3.1.1. Therefore, we may consider a knowledge graph as a subset of an original knowledge base, focused on the concepts pertaining to an input phrase. Fig. 1 shows a knowledge graph example.

Knowledge graphs have been used in natural language processing tasks for a long time, such as network text analysis (Popping, 2003), semantic relatedness (Navigli and Ponzetto, 2012b), word sense disambiguation (Moro et al., 2014), semantic parsing (Heck et al., 2013), sentiment analysis (Franco-Salvador et al., 2015) and in cross-language scenarios such as plagiarism detection (Franco-Salvador et al., 2016), and document categorization (Franco-Salvador et al., 2014).

In contrast to previous works in automated phrase sampling methods for text entry (Leiva and Sanchis-Trilles, 2014; Paek and Hsu, 2011), which have to learn features for each new supported language or domain, we aim at standardizing the knowledge representation of the "phrase universe" (i.e., a given language and/or domain). Therefore, we selected BabelNet[2] (Navigli and Ponzetto, 2012a), the multilingual semantic network with the widest concept and language coverage, as a common knowledge base to generate our knowledge graphs. Before delving into the details of knowledge graph generation, we will describe BabelNet and introduce its basic terminology below.

#### 3.1.1. BabelNet

BabelNet is a multilingual semantic network, comprising lexicographic and encyclopedic knowledge in different languages. BabelNet connects concepts and named entities in a very large network of semantic relations. These relations are established from about 14 million entries, known as Babel synsets. Each synset represents a given sense/meaning and agglutinates all the synonyms that express such meaning in a range of different languages. For example, the "house" concept (as in "building") has the same synset when querying BabelNet with "casa" (Spanish) or "ДОМ" (Russian).

In BabelNet, synsets and relations are obtained from the automatic mapping onto WordNet of Wikipedia,[3] OmegaWiki,[4] Wiktionary,[5] Wikidata,[6] and Open Multilingual WordNet,[7] among others. The available syntactic categories are: noun, verb, adverb, and adjective. The full list of resources and statistics is available at http://babelnet.org/stats. The current version of BabelNet (3.7) includes 13,801,844 synsets, covering 745,859,932 senses across 271 languages.

#### 3.1.2. Knowledge graph generation

We follow the approach of Navigli and Lapata (2010) to create our knowledge graphs, which is a four step-approach, described as follows.

*(i) Part-of-speech tagging and lemmatization.* Initially we process a phrase $x$ with tokenization, multi-word extraction, part-of-speech (POS) tagging, and lemmatization. For POS tagging we use the Stanford CoreNLP toolkit[8] and TreeTagger[9] as fallback for those languages not supported by CoreNLP. For multi-word extraction, we implemented our own tool based on pattern matching, so that common POS tag sequences involving nouns (e.g. Noun + Noun, Adjective + Noun, Noun + Preposition + Noun, ...) that exist in BabelNet are collapsed into a single Noun POS tag. For

[2] http://babelnet.org

[3] http://wikipedia.org
[4] http://omegawiki.org
[5] http://wiktionary.org
[6] http://wikidata.org
[7] http://compling.hss.ntu.edu.sg/omw/
[8] http://stanfordnlp.github.io/CoreNLP/
[9] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

example, the multi-word "United States" (Adjective + Noun) is eventually tagged as Noun, since the "United States" concept exists in BabelNet.

As a result, we obtain a set of tuples $\mathcal{T} = \{t_1 \cdots t_i\} : t =$<lemma/tag> for each input phrase. For example, the phrase '`Mary had a little lamb`' produces $\mathcal{T}$ ={ Mary/NNP, have/VBD, a/DT, little/NN, lamb/NN }), where NNP denotes a proper noun, VBD denotes a verb (lemmatized in the infinitive form), DT denotes a determiner, and NN denotes a common noun. POS tags not available in BabelNet are discarded when creating the knowledge graph.

*(ii) Populating the graph with initial concepts.* Next, we create an initially-empty knowledge graph $G = (\mathcal{V}, \mathcal{E})$ such that $\mathcal{V} = \mathcal{E} = \emptyset$. We populate the vertex set $\mathcal{V}$ with the set $S$ of all the synsets in BabelNet comprising any tuple $t$ in the language $L$, i.e.

$$S = \bigcup_{t \in \mathcal{T}} \text{Synsets}_L(t) \tag{1}$$

where $\text{Synsets}_L(t)$ is the set of synsets that contains a tuple $t$ in the language $L$.

*(iii) Creating the knowledge graph.* Then we create the knowledge graph by searching in BabelNet the set of paths $\mathcal{P}$ connecting pairs of synsets in $\mathcal{V}$. Formally, for each pair $(v, v') \in V$ such that $v$ and $v'$ do not share any lexicalization in $\mathcal{T}$, for each path in BabelNet $v \rightarrow v_1 \rightarrow \cdots \rightarrow v_n \rightarrow v'$, we update $\mathcal{V} := \mathcal{V} \cup \{v_1 \cdots v_n\}$ and $\mathcal{E} := \mathcal{E} \cup \{(v, v_1) \cdots (v_n, v')\}$. That is, we add all the path vertices and edges to $G$.

Following the approach of Navigli and Ponzetto (2012a), the path length is bounded to a maximum length of 3, which means that at most 2 intermediate and potentially new concepts (not present in the input phrase) will be considered, in order to avoid an excessive semantic drift. As a result of populating the graph with intermediate edges and vertices, we obtain a knowledge graph which models the information available in the input phrase $x$.

*(iv) Knowledge graph weighting.* The last step consists of weighting all the concepts and semantic relations of the knowledge graph $G$. For weighting relations we use the original weights from BabelNet, which estimate the degree of relatedness between two synsets as a function of the overlap between their gloss words and synset neighbors. At this point, we remove the edges below a certain threshold that represents a low semantic relationship; for which we use the threshold provided by Navigli and Ponzetto (2012a).

For weighting concepts, different methods have been tested in the past, including for example the PageRank algorithm (Page et al., 1998). In this work, we score each concept using its own outdegree feature, which has proved to achieve the best results (Navigli and Ponzetto, 2012a). Finally, we normalize the weighted concepts as a function of the sum of the outgoing relations.

### 3.2. Phrase sampling

We approach phrase sampling as a graph-based ranking task. Having some set of phrases as input, typically a text corpus that can be obtained from several public resources,[10] we should rank first those phrases that meet some quality criteria, in order to determine the best candidates for text entry experiments. The key question is therefore which quality criteria should we aim for ranking.

Traditional readability tests like Coleman-Liau index (Coleman and Liau, 1975) or ARI (Senter and Smith, 1967) were designed to gauge the understandability of whole documents (not single phrases), so they do not suit our needs. Further, while readability formulas are considered to be predictors of reading ease, they do not evaluate how well a reader will understand the ideas in the text (Klare, 1976). Further, readability tests are overly simplistic and systematically retrieve extremely short phrases (Leiva and Sanchis-Trilles, 2014). Similarly, computational models such as Flesch (Flesch, 1948), FOG (Gunning, 1952), or SMOG (McLaughlin, 1969) are also available for predicting the readability of texts. Although these methods assume long, well written text, and only extract basic lexical features like average number of characters per word (Kanungo and Orr, 2009).

In this paper, we followed previous works that suggested using the character error rate (CER) as a proxy of memorability (Kristensson and Vertanen, 2012; Leiva and Sanchis-Trilles, 2014; MacKenzie and Soukoreff, 2002; Vertanen and Kristensson, 2011b). Furthermore, it has been shown that CER correlates with typing effort, which can be seen as an indirect measure of cognitive processing. Then, we will focus on modeling CER so that phrases with the lowest predicted CER are ranked first. Formally, given an unordered large set of phrases $\mathcal{X} = \{x_1 \cdots x_n\}$ the goal of sampling is to obtain a smaller phrase set $\mathcal{R} \subset \mathcal{X}$ of size $|\mathcal{R}| = N \ll |\mathcal{X}|$ that satisfies the $k$th order statistic, i.e., the phrase at the $k$th position in the ranking satisfies

$$\mathcal{R}_{(k)} = \arg\min_{\mathcal{X}} \text{CER}(\mathcal{X}) \tag{2}$$

where $\text{CER}(\mathcal{X})$ is a CER-based scoring function over $\mathcal{X}$, to be described later.

Notice that the MemRep technique (Leiva and Sanchis-Trilles, 2014) approached phrase sampling as a ranking task, but it required as input an additional large dataset to model the language of interest. Kaps removes this need, simplifying thus the sampling procedure to a great extent. This way, the experimenter just has to decide how many phrases she need to conduct the text entry experiment (e.g. N=2,000 phrases) and Kaps will return the top-$N$ phrases from the ranking.

In this work, we model CER as a function of language-independent lexical and semantic features. We extract a number of features from the knowledge graphs based on their topology and sparsity (Mihalcea and Radev, 2011), grammatical categories, syntactics, etc. (see the complete feature list in Appendix B). Then, we perform a feature selection algorithm, in order to keep the best subset of phrase-level features. Feature selection is necessary because some features might be redundant and/or dominate the others. In general, selecting the right features can mean a difference between low performance with long computation times and high performance with short computation times. In the following we describe how feature selection was used to derive our model.

### 3.3. Model estimation

The CER model was fitted according to an iterative re-weighted generalized linear regression (Nelder and Weddrburn, 1972), which derives maximum likelihood estimates of the model parameters without expecting any particular distribution of such parameters.

As in other feature engineering experiments, it is expected that not all features considered will be strong model predictors. So, we used the Bayesian Information Criterion (BIC) approach (Schwarz, 1978) to remove the non-significant features, statistically wise. The BIC criterion minimizes the information loss of the fit:

$$\text{BIC}(k) = -2\log(\mathcal{L}) + k\log(n) \tag{3}$$

where $\mathcal{L}$ is the maximum likelihood for the generalized linear model, $k$ is the number of parameters in the model, and $n$ is the number of observations (data points) in the dataset.

BIC is an interesting choice to fit model parameters because it tends to build simple models and converges as the number of observations increases (Burnham and Anderson, 2004). BIC asymptotically attempts to estimate the 'true' model in a stepwise procedure, using iteratively re-weighted least squares. It is based on the idea that minimizing the

[10] For example: DataHub (https://datahub.io/dataset), ELRA (http://catalog.elra.info), LDC (https://www.ldc.upenn.edu), UCI (http://archive.ics.uci.edu/ml/), or the Europarl datasets (http://www.statmt.org/europarl/).

relative distance between the (unknown) 'true' model and the tentative model yields the optimal model.

The CER model was trained with the EnronMobile dataset (Vertanen and Kristensson, 2011b), which provides 2.2K unique sentences derived from a memorization experiment. BIC revealed that a subset of 10 features (out of 78, see Appendix B) were statistically significant predictors. The final model yielded a good fit (Adj. $R^2 = 0.65$),[11] with the following combination of features, in descending order of statistical relevance ($p$-value):

$$\begin{aligned}\mathrm{CER}' = {} & 1.03 \cdot \mathbf{Nw} + 0.72 \cdot \mathbf{Achr} - 4.41 \cdot \mathbf{PwWvS} + 6.66 \cdot \mathbf{Pn} - 4.97 \cdot \mathbf{SDo} \\ & + 0.21 \cdot \mathbf{Mo} + 4.58 \cdot \mathbf{PwWs} + 8.13 \cdot \mathbf{Aio} - 0.05 \cdot \mathbf{Mi} + 1.03 \cdot \mathbf{SDi}\end{aligned} \tag{4}$$

where the acronyms correspond with:

- CER′: Predicted character error rate.[12]
- **Nw**: Number of words (i.e., phrase length).
- **Achr**: Average number of characters per word.
- **PwWvS**: Percentage of synsets that are verbs, relative to phrase length.
- **Pn**: Percentage of words that include numbers, relative to phrase length.
- **SDo**: Standard deviation of the concept outdegree.
- **Mo**: Maximum concept outdegree.
- **PwWs**: Percentage of words with synsets, relative to phrase length.
- **Aio**: Average indegree and outdegree of the concepts.
- **Mi**: Maximum concept indegree.
- **SDi**: Standard deviation of the concept indegree.

Interestingly, the model picks two lexical features as in MemRep: number of words (**Nw**) and average number of characters per word (**Achr**). This reveals that short phrases comprising short words will ensure memorability. However, we should stress the fact that phrases with a few words or very short words would not be appropriate for conducting a text entry experiment, as they will rarely be representative either of general language or a particular task or domain; cf. the discussion at the beginning of this section.

The model suggests that short texts containing a few close concepts (about the same topic overall, with concepts related to each other) require less cognitive processing and, as a consequence, are more memorable (Sweller, 1994). Therefore, the higher the percentage of words with synsets (**PwWs**), the more diverse concepts is handling the phrase. As a result, the phrase will be more difficult to assimilate in a first read. Similarly, high average values of indegree and outdegree (**Aio**) or maximum outdegree (**Mo**) indicate that the graph representing the phrase is large and more connected, typically because of a larger phrase length and a high number of concepts handled. This harms memorability and may also compromise representativeness, by introducing too much (probably noisy, unnecessary, or out-of-topic) information. The positive weight of the standard deviation of the indegree (**SDo**) follows the same logic: it is better to have all the concepts related with a reduced number of central concepts (representing the main topic of the phrase), instead of being all of them connected with very diverse concepts.

According to the final model, lexical features such as the percentage of words with numbers (**Pn**) or average number of characters per word (**Achr**) are related to the complexity of the text, which affects both memorability and typing effort. In consequence, the higher these values the harder to process the phrase. In contrast, phrases will be easier to transcribe from memory if they have a central and highly connected component. This is indicated by the high values of maximum indegree and outdegree standard deviations (**SDo, SDi**). Higher values of the latter case are observed in graphs with high outdegree variability. This variability seems beneficial for ensuring memorable, representative, and semantically correct phrases: it indicates that there are connected components, some of them with high connectivity, i.e., high semantic relatedness between concepts, but the graph is not so dense as to be representing a very complex phrase, which would be more difficult to assimilate.

Finally, the model suggests that the presence of verbs (units of language that represent actions) is beneficial to reduce cognitive processing effort. This is corroborated by previous work on cognitive psychology (Rayner and Duffy, 1986). The influence of verbs in the model is represented by the ratio of synsets that are verbs in the input phrase (**PwWvS**). Compared to other linguistic categories such as nouns or adjectives, verbs appear less often in the same phrase. Therefore, phrases that contain a higher percentage of verbs are easier to understand, process, and memorize; in comparison to other categories with more potential vocabulary and lexical complexity, cf. '`hearings are planned into last summer's stubborn wildfires`' vs.'`we were dancing, screaming and enjoying all night`' (both phrases have the same number of words).

It should be noted that some phrases in the input set $\mathcal{X}$ may contain no semantic information, in which case they would be ranked on the basis of the lexical features

(**Nw, Achr, Pn**). Therefore, phrases with semantic information are always ranked first, i.e., those phrases having at least one feature of semantic nature (**Aio, SDi, Mo, Mi, PwWs, PwWvS, SDo**).

## 4. Evaluation

We conduct five experiments to validate the capabilities of Kaps, which we compare against their closest peers. First, we conduct a small-scale evaluation involving 20 users and 2 languages, and show that using English phrases with *non-native* English speakers results in higher memorization times and more transcription errors. Second, we show that Kaps is a competitive memorability predictor, achieving similar performance to a state-of-the-art, more complex classifier. Third, we show that Kaps ensures representativeness, providing phrase sets that are highly correlated with the population from which phrases were sampled. Fourth, we show that Kaps is semantically accurate, providing the same performance regardless the available language coverage or the number of sampled phrases. Finally, we conduct a large-scale evaluation involving 200 *native* speakers in 10 languages, and show that Kaps is on par with MemRep , which is a more sophisticated phrase sampling method for text entry.

### 4.1. Language proficiency analysis

Most text entry evaluations are conducted with English-only datasets (cf. Table 1), so it is unclear what is the impact of the phrase set on participants' language proficiency. This study is important because the internal validity of text entry experiments could be compromised, by inflating artificially e.g. text entry throughput. In fact, previous text entry experiments that have examined the impact of phrase set on language proficiency have focused on text entry speed rates, measured in words per minute (Isokoski and Linden, 2004; Lyons and Clawson, 2012).

#### 4.1.1. Data

We compare Kaps against its closest peer MemRep , for which we replicate the experimental conditions of the MemRep study conducted by Leiva and Sanchis-Trilles (2014). In that study, phrases from the Europarl dataset (1.8 million sentences from the proceedings of the European parliament) were used as input stimuli both in English and Spanish, resulting in two MemRep datasets of 500 phrases each. These datasets

[11] Contrary to $R^2$, the Adjusted $R^2$ explains the model variation by only the independent variables (i.e., the model features) that actually affect the dependent variable (in this case, CER). In other words, if we keep adding features to the model $R^2$ will always increase, which is misleading, but Adj. $R^2$ will not increase.

[12] We add the prime symbol to avoid confusion with the actual character error rate committed by the users.

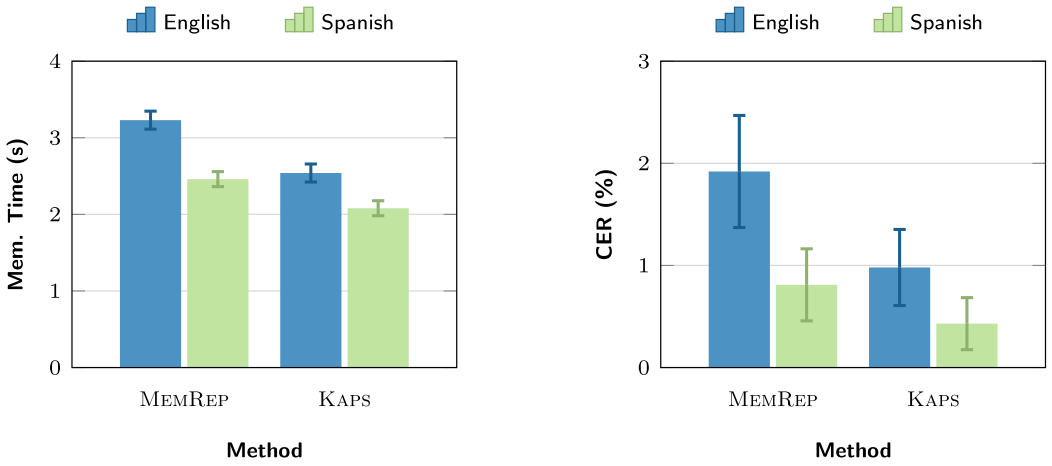


**Fig. 2.** Results of the language proficiency analysis. Error bars denote 95% CIs.

are publicly available at https://luis.leiva.name/memrep/. Therefore, we sampled the same number of phrases from the same dataset in the same languages with Kaps . Phrases were lowercased and punctuations symbols were removed, as in the previous MemRep study.

#### 4.1.2. Participants

We recruited 20 native Spanish speakers (8 female) aged 26–37 using our University's mailing lists. Participants had to enter phrases both in their native language and in English. All participants had a qualified intermediate or advanced English degree according to the Common European Framework of Reference for Languages.[13] Each participant was paid a 5 Euro gift voucher.

#### 4.1.3. Method

Both the participants' recruitment and the evaluation procedure were the same as in the previous MemRep study: each participant started with phrases in one language (either English or Spanish) and then entered phrases in the other language. The language order was randomized, ensuring that half of the users would start with the English phrases and the other half would start with the Spanish phrases. Within each language, users were randomly presented with Kaps and MemRep phrases (they were not told which were which), ensuring that half of the phrases were MemRep phrases and the other half were Kaps phrases.

Each participant was shown a phrase for 5 seconds or until the first keystroke, whatever happened first. Afterward, the phrase disappeared and users had to write it (as much as they could remember) with a physical keyboard, be it QWERTY, QWERTZ (mostly used by German speakers), or AZERTY (mostly used by French speakers). By the end of the study, each participant entered 50 phrases from both datasets, resulting in 20 participants × 50 phrases × 2 sampling techniques = 2,000 annotated phrases in total.

#### 4.1.4. Results

The results of this experiment are show in Fig. 2. We report the actual character error rate (CER, in %), which is the standard measure of text entry accuracy. CER is computed as the Damerau-Levenshtein distance between the text transcribed by the user and the reference text, normalized by the number of characters in the reference text. We also report the time since the phrase was loaded until the very first keypress (in seconds). This measure is an estimate of the time spent memorizing each phrase. As previously stated, the maximum memorization time allowed would be 5 s per phrase, since after that time the phrase would have disappeared.

As can be observed in Fig. 2, participants spent more time memorizing the English phrases than the Spanish ones. Participants also committed more errors when entering the English phrases. Kaps performed better than MemRep in both languages.

We conducted a repeated measures two-way ANOVA as omnibus test, with language and sampling method as factors. We assessed the differences in memorization time and CER. The sampling method was a within-subject factor, whereas language was a between-subject factor; thus it is a mixed design. And since it is also a repeated measures design, we controlled for natural variation from participant to participant.

We observed a significant effect attributed to the language of the phrases: memorization time $\left[F_{(1,57)} = 16.47, p < .001, \eta_p^2 = 0.29\right]$; CER $\left[F_{(1,57)} = 10.45, p < .01, \eta_p^2 = 0.18\right]$. We also observed a significant effect attributed to the sampling method: memorization $\left[F_{(1,57)} = 12.63, p < .001, \eta_p^2 = 0.22\right]$; CER $\left[F_{(1,57)} = 5.67, p < .05, \eta_p^2 = 0.10\right]$. Effect sizes suggest small to moderate practical importance. No significant interaction between language and sampling method were found. No pairwise comparisons were performed because each factor has only two levels. We concluded that English phrases are harder to memorize and more error-prone for Spanish speakers, and that Kaps outperforms MemRep in this task.

It was expected that participants would perform better if the phrases were shown in their native language. What was surprising, though, is the fact that English MemRep phrases took almost 1 s more on average to memorize in comparison to their Spanish counterparts, considering that all participants had an official English certification. On the contrary, Kaps phrases performed similarly in both languages (0.4 s of difference). These results suggest that Kaps produces phrases that are easier to memorize than MemRep for non-native English speakers. We should mention that both sampling methods produced phrases of very similar length, around 4 words on average.

Regarding CER, it can be observed that English phrases had twice higher values in comparison to their Spanish counterparts, and this was true for both sampling methods. Again, this is explained because English was not the native language of the participants. Further, participants committed significantly more errors while entering the English MemRep phrases. These results suggest that Kaps produces phrases that are easier to type than MemRep , when participants are not shown phrases in their native language. In sum, we conclude that phrases should be shown in the native language of the study participants, otherwise performance will be artificially inflated. This experiment also establishes Kaps as a very competitive automated phrase sampling method for non-English native speakers.

[13] http://www.coe.int/lang-CEFR

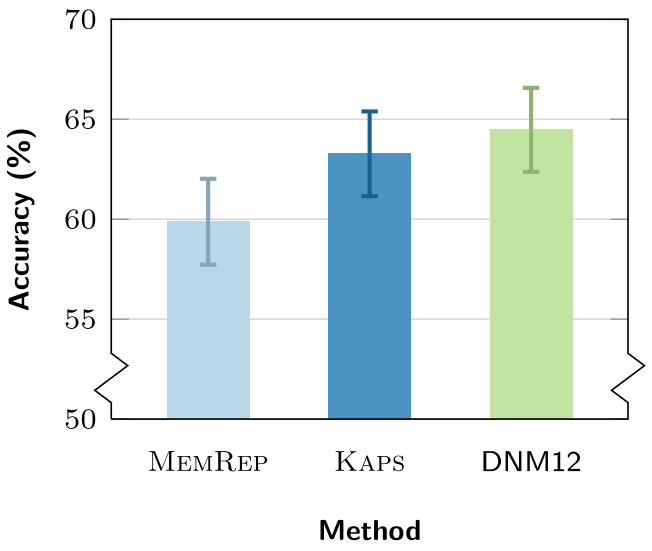


**Fig. 3.** Results of the memorability analysis. Error bars denote 95% CIs.

### 4.2. Memorability analysis

In order to evaluate the memorability capabilities of Kaps from a theoretical perspective, we conducted a classification experiment. Memorability is the degree to which something is remembered or forgotten. As simple as it sounds, defining what makes an phrase memorable is subtle; see Section 2. However, in the context of this experiment we will stick to the definition above (memorable = easy to remember).

#### 4.2.1. Data

We selected the Cornell Movie Quotes (MovieQuotes) dataset (Danescu-Niculescu-Mizil et al., 2012) as gold standard, since it is annotated with memorability information and has been used by previous researchers. Notice that we cannot use the EnronMobile dataset in this experiment, since testing our model with the same data used for training would give it an unfair advantage over other methods. In addition, we should note that MemRep was also trained with the EnronMobile dataset.

The MovieQuotes dataset was constructed from roughly 1,000 movies, varying in genre, era, and popularity; resulting in 2,200 sentence pairs, each pair comprising a memorable and a non-memorable movie quote. The dataset is publicly available at http://www.cs.cornell.edu/~cristian/memorability.html. Each sentence pair comprises transcriptions of lines spoken by the same actor in the same movie. Moreover, both sentences in the same pair have the same length and appear as nearly as possible in the same movie scene.

#### 4.2.2. Method

We formulate a pairwise comparison task: given a sentence pair (two movie quotes), determine which sentence is the memorable one. We use different methods to classify each sentence pair and measure the classification accuracy. Concretely, we compare the results of Kaps against MemRep as well as the predictive method devised by the MovieQuotes authors (Danescu-Niculescu-Mizil et al., 2012), which we will refer to as DNM12 for short, as baseline condition. This method is a Support Vector Machine (SVM) classifier that employs a combination of 52 features based on the distinctiveness and generality of a movie quote.

#### 4.2.3. Results

The results of this experiment are shown in Fig. 3. For the null hypothesis of random guessing, which would achieve 50% of accuracy, all results are statistically significant ($p < .001$). We used a $\chi^2$ test to test the alternative hypothesis that "not all proportions are equal". The test revealed statistical significance $\left[\chi^2_{(2,N=2200)} = 10.69, p < .01, \phi = 0.07\right]$, which means that at least one proportion is different from the others. The effect size suggests small practical importance.

Therefore, to identify where the differences between proportions exist, we performed pairwise comparisons (two-tailed $t$-tests) as post-hoc test of significance. The test used the Bonferroni correction, to reduce the chances of obtaining false positives (Type I errors). We observed that DNM12 performed statistically better than MemRep ($p < .01$) but no better than Kaps ($p > .05$, n.s.) The difference between MemRep and Kaps was found to be statistically significant ($p < .05$).

As observed, MemRep fared worse than its peers. This can be explained because MemRep 's language model (learned from Europarl) is deemed as insufficient for this classification task, resulting in phrases that underestimate the memorability classification results.

On the other hand, DNM12 and Kaps performed equally similar. We should note that DNM12 is a fairly complex memorability classifier, trained for the specific task of classifying memorable movie quotes. For example, it requires a training partition to build a SVM classifier using a 10-fold cross-validation procedure, which means that it must learn to classify each task, one dataset at a time. In contrast, both MemRep and Kaps are more general approaches in this regard: MemRep uses a large language model (only has to be learned once) and Kaps relies on a multilingual semantic network (no language modeling is required).

In light of the results observed in this experiment, we can see that Kaps captures the memorability subtleties described by a state-of-the-art classifier (Danescu-Niculescu-Mizil et al., 2012), even though considering that representativeness might have an opposite effect on memorability (Leiva and Sanchis-Trilles, 2014). It is by incorporating phrase semantics that we can predict memorability more successfully, even achieving state-of-the-art performance. We conclude that Kaps is a very competitive approach to ensure phrase memorability.

### 4.3. Representativeness analysis

Representativeness is the closeness of characteristics of a sample to the corresponding characteristics of the population from which the sample has been taken (Kruskal and Mosteller, 1979). A judgment as to the degree of representativeness of a sample can be made for example, by comparing the characteristics that have been measured in both the sample and the population. Obviously, when the sample size increases such sample becomes more representative, since it approaches the population size. Eventually, any indicator or set of characteristics measured over the sample will correlate perfectly when the entire population has been examined. Therefore, a good representative sample should correlate highly with their population even for small sample sizes.

#### 4.3.1. Data

We conducted this experiment over the datasets analyzed so far: Europarl in English (EN) and Spanish (ES), and the MovieQuotes dataset.

#### 4.3.2. Method

In this experiment, we compute a number of phrase-level features that define a phrase population. Then, we report the Pearson's correlation coefficient to judge the degree of representativeness of different sample sizes. We express the phrase set sizes as a percentage of the phrase population, in order to compare better the different datasets analyzed (to be discussed later), since each dataset has a different number of phrases.

We compare Kaps against the state-of-the-art automated phrase sampling method (MemRep) and random selection as baseline condition (averaged over 10 folds to reduce any possible bias).

We use the following phrase-level features to measure phrase representativeness:

**No. words** Number of words in the phrase (i.e., phrase length).

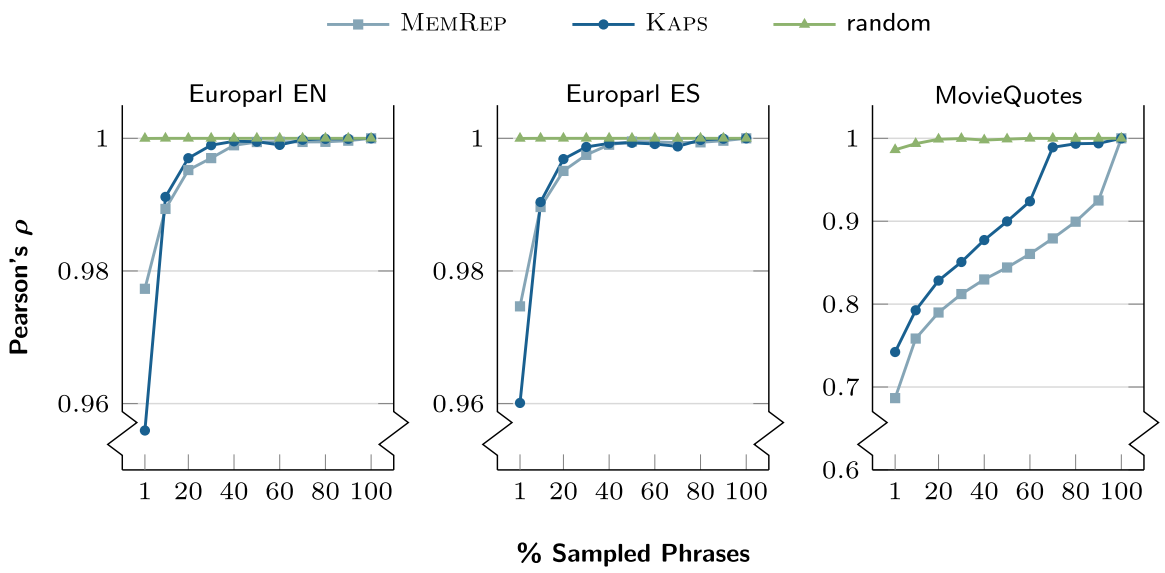


**Fig. 4.** Results of the representativeness analysis.

**No. chars** Average number of characters per word.
**SD chars** Standard deviation of the number of characters per word.
**OOV ratio** Number of infrequent words, relative to phrase length.
**No. classes** Number of syntactic categories (noun, verb, adverb, and adjective) found in the phrase.
**Synsets ratio** Number of synsets found in the phrase, relative to phrase length.

The first four measures are derived from previous work (Leiva and Sanchis-Trilles, 2014; Sanchis-Trilles and Leiva, 2014), whereas the last two measures are our own, which are particularly relevant to account for phrase semantics.

We should remark that MemRep and Kaps use different phrase features for sampling, and we use the list described above as a common criteria to compare the phrase samples provided by each method. Such list is general enough as not to bias the results toward a particular sampling method: **No. words** and **No. chars** are common to MemRep and Kaps , **SD chars** and **OOV ratio** are MemRep -specific, and **No. classes** and **Synsets ratio** are Kaps -specific.

It should be noticed that **OOV ratio** in Kaps was computed with respect to the most frequent vocabulary according to Wikipedia (see Table 3). Also, when computing the **No. classes** and **Synsets ratio** we had to generate the corresponding knowledge graphs for MemRep phrases.

#### 4.3.3. Results

As predicted, the correlation between the sampled phrases and their population increases as the number of the sampled phrases increases. However, it is interesting to notice *how* such increment is produced. Fig. 4 shows the results. As can be observed in the figure, there is a high correlation even with very small sample sizes (less than 20% of the phrase population). This is especially true with regard to the Europarl datasets, which was somehow unsurprising for MemRep since it ensures representativeness by design. What was surprising, however, is the fact that Kaps outperforms MemRep for all sample sizes higher than 1% of the population. This is especially true with regard to the MovieQuotes dataset. In general, an elbow-shape curve is preferred, something that both methods achieve.

As observed, random sampling correlates the best for any sample size. This result is unsurprising, since this selection method is unbiased by definition, so on average any sample would represent the population more accurately than any other sampling method that applies some optimization criteria, like MemRep and Kaps do. Nevertheless, random sampling does not guarantee an adequate selection of phrases for conducting text entry evaluations, see Section 2.2.

We used the $z$-test (two-tailed, with Fisher transformation) to analyze the differences between MemRep and Kaps correlations. We applied the Bonferroni correction to guard against over-testing the data. For samples of 1% of the population size, Kaps performed better than MemRep in Europarl EN $[z = 10.02, p < .001]$, but in Europarl ES MemRep performed better than Kaps $[z = 6.85, p < .001]$. For samples of 10% of the population size, Kaps performed better than MemRep in both Europarl datasets: EN $[z = -8.81, p < .001]$; ES $[z = -3.56, p < .001]$. Differences were not statistically significant for sample sizes higher than 40% of both datasets.

As regarding the MovieQuotes dataset, the differences between MemRep and Kaps were not found to be statistically significant for sample sizes below 10% $[z = -0.84, p > .05, \text{n.s.}]$. It was for sample sizes between 20% $[z = -2.23, p < .05]$ and 90% $[z = -38.34, p < .001]$ where differences were found to be statistically significant. We conclude that Kaps produces phrase samples that are representative of the population where phrases come from, no matter the sample size, and that it outperforms MemRep for small sample sizes and above.

### 4.4. Semantic coverage analysis

In this section we analyze the performance of Kaps as a function of the semantic coverage provided by the knowledge base employed (BabelNet). By definition, synsets are language-independent, therefore Kaps should perform similarly in other languages. We conducted two experiments to test this idea.

In both experiments, we analyze performance in terms of phrase memorability and phrase representativeness in 10 different languages. The choice of languages was based on their popularity according to Wikipedia, as noted in a previous study by Sanchis-Trilles and Leiva (2014) which will be described in the next section.

#### 4.4.1. Data and method

In the first experiment, we measure memorability by classifying the phrases of the MovieQuotes dataset using only the synsets available in a particular language, one language at a time. We use the MovieQuotes dataset since it is labeled at the phrase level. Again, we should mention that we cannot use the EnronMobile dataset here, since it would not be fair testing our method with the same data used for training. Also notice that the MovieQuotes dataset is only available in English, so

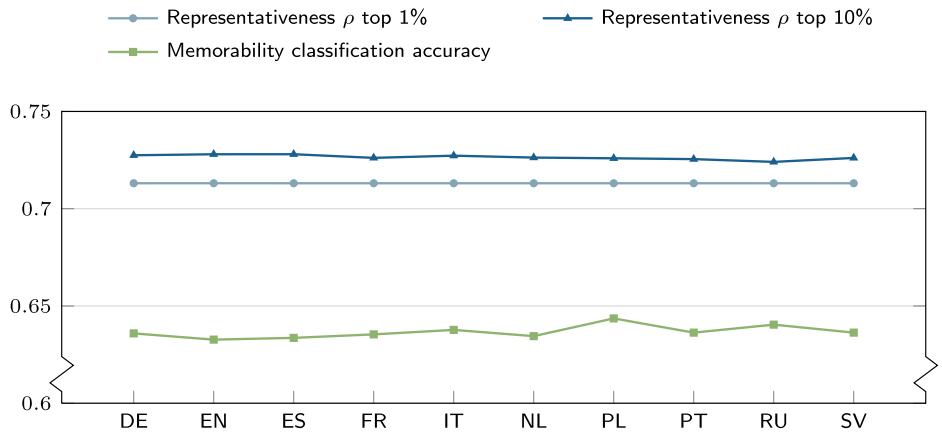


**Fig. 5.** Semantic coverage results in terms of phrase memorability and phrase representativeness across ten different languages, denoted with their ISO 639-1 language codes; see also Table 2.

**Table 2**
Number of synsets per language. Source: BabelNet.

| Language | ISO 639-1 Code | No. Synsets |
|---|---|---|
| German | DE | 1.37 M |
| English | EN | 3.90 M |
| Spanish | ES | 1.04 M |
| French | FR | 1.33 M |
| Italian | IT | 1.05 M |
| Dutch | NL | 1.20 M |
| Polish | PL | 0.98 M |
| Portuguese | PT | 0.84 M |
| Russian | RU | 0.98 M |
| Swedish | SV | 0.64 M |

the purpose of the first experiment is determining whether we could derive semantic information from other languages to supplement the English knowledge.[14] Therefore, this analysis aims for proving whether the synsets' linguistic independence allows us to achieve this goal.

In the second experiment, we measure representativeness by correlating phrase sets of size 1% and 10% of the input phrase population with the total phrase population. Therefore, this analysis aims for assessing the effect of knowledge base coverage on different phrase set sizes.

In addition, we perform a deeper analysis with a well-supported language such as English and a lesser-supported language such as Swedish (Table 2). In fact, according to BabelNet, English is the best-supported language whereas Swedish is the least-supported one among the 10 languages selected for analysis. For these particular languages, we measure such representativeness correlation as a function of the number of synsets available, from no semantic information (0% of synsets are employed) to full semantic information (100% of synsets are employed). Therefore, this analysis aims for proving whether synsets allow KAPS to achieve better performance than using no semantic information.

#### 4.4.2. Results

Fig. 5 shows the results achieved by KAPS when using exclusively the synsets of the languages under study, one language at a time. These results indicate that even with languages having a low number of synsets (cf. Swedish), KAPS is able to exploit the semantic information available in BabelNet to obtain competitive results pertaining both memorability and representativeness.

[14] We report results for the English language as a baseline condition.

Compared to the original results with the English language, we note that using a less massive but robust synset inventory (cf. Polish), KAPS achieves the same performance levels in terms of representativeness. This is so because, as expected, synset are language-independent; and indeed they provide an interesting concept abstraction (see Section 3.1.1) that can be used in other languages.

Notice that some languages such as Russian or Polish have a smaller synset inventory than English but nonetheless they achieve a slightly higher accuracy. This is explained by the fact that their synset inventory is mainly based on the Wikipedia information available for these languages. Such information is more reduced but covers the essential encyclopedic human knowledge, which results in more robust synsets than those of English.

Figure 6 shows the results achieved by KAPS as a function of the percentage of available synsets in BabelNet. For a particular percentage value, we randomly select the synsets and average the results over 10 folds to reduce a possible bias of such random selection. As can be observed in the figure, even a low percentage of semantic information leads KAPS to achieve better performance over its non-semantic variant (0% of synsets).

Overall, KAPS improves when the synset coverage increases, as predicted, but it is worth mentioning that it also achieves competitive results with a low synset coverage. These results are consistent both with languages having a low number of synsets (Fig. 6b) and languages having a high number of synsets (Fig. 6a). This puts forward the fact that BabelNet, the largest multilingual semantic network, is an excellent knowledge base that can be successfully exploited as a common resource for phrase sampling.

### 4.5. Text entry performance with native speakers

The last experiment comprises a multilingual evaluation with *native* speakers, following a crowsourcing scenario and comprising the 10 languages analyzed so far. Aiming at a general study, we used the traditional QWERTY keyboard as a common input technique. Further, being a crowdsourced user study, we needed to ensure that all participants used the same input device.

#### 4.5.1. Data

Since MEMREP requires an additional text corpus as input to model the knowledge of a language, we used the Wikipedia datasets provided by Sanchis-Trilles and Leiva (2014) in the 10 languages of interest: German, English, Spanish, French, Italian, Dutch, Polish, Portuguese, Russian, and Swedish. These datasets comprises the Wikipedia articles that had at least 1 million articles as of 2014. According to Sanchis-Trilles and Leiva (2014), this language selection attempted to "comprise

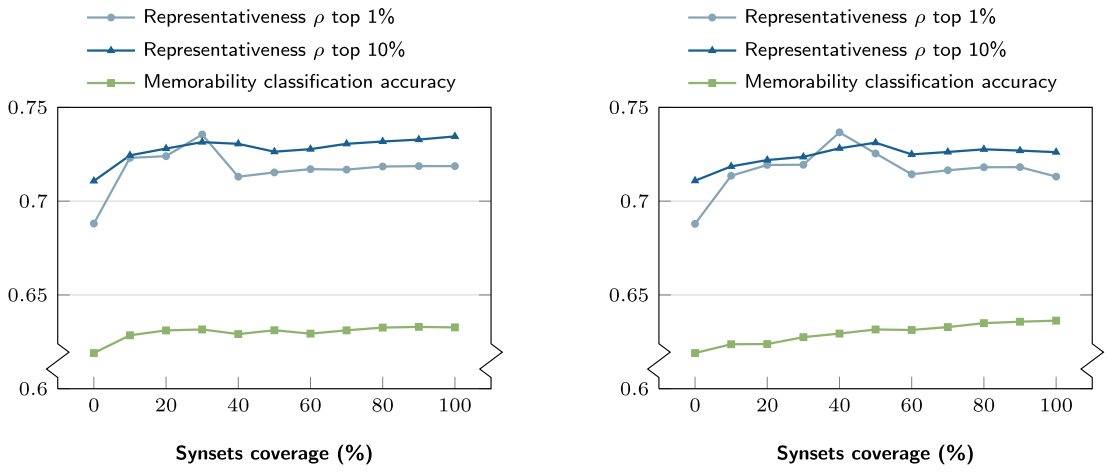


**Fig. 6.** Semantic coverage results in terms of synsets availability for the best-equipped language (a) and the least-equipped language (b) among the 10 selected languages.

**Table 3**
Overview of the Wikipedia corpus, required by MemRep to build its language models.

| Language | Sentences | Words | Vocabulary | Size |
|---|---|---|---|---|
| German | 16.0 M | 534 M | 7.3 M | 3.6GB |
| English | 63.2 M | 1,232 M | 5.8 M | 6.9GB |
| Spanish | 10.1 M | 374 M | 2.6 M | 2.2GB |
| French | 13.6 M | 428 M | 2.7 M | 2.5GB |
| Italian | 9.2 M | 300 M | 2.3 M | 1.8GB |
| Dutch | 7.7 M | 192 M | 2.6 M | 1.2GB |
| Polish | 6.1 M | 156 M | 2.5 M | 1.1GB |
| Portuguese | 5.2 M | 165 M | 1.7 M | 4.1GB |
| Russian | 6.9 M | 20 M | 4.1 M | 2.7GB |
| Swedish | 5.5 M | 114 M | 2.4 M | 0.8GB |

**Table 4**
Overview of the OpenSubtitles dataset for all the languages under study.

| Language | Sentences | Words | Vocabulary | Size |
|---|---|---|---|---|
| German | 25.1 M | 137 M | 611.4 k | 0.8GB |
| English | 265.1 M | 1,575 M | 812.3 k | 14.0GB |
| Spanish | 212.1 M | 1,219 M | 1,153.2 k | 6.8GB |
| French | 129.8 M | 781 M | 694.2 k | 3.9GB |
| Italian | 65.7 M | 376 M | 610.1 k | 2.0GB |
| Dutch | 127.8 M | 737 M | 1,018.8 k | 3.7GB |
| Polish | 183.9 M | 843 M | 1,113.5 k | 7.3GB |
| Portuguese | 141.0 M | 764 M | 684.9 k | 1.0GB |
| Russian | 28.9 M | 144 M | 962.3 k | 1.5GB |
| Swedish | 38.1 M | 200 M | 660.2 k | 1.1GB |

a big part of the languages that might be subject of study in text entry experiments". However, these datasets were never evaluated with actual users. Table 3 summarizes these datasets.

We used Kaps to sample 2,000 phrases from the OpenSubtitles dataset in the same 10 languages as in the Wikipedia datasets (20,000 sampled phrases in total). Table 4 summarizes these datasets. Some examples of the resulting sampled phrases are shown in Appendix A.

As previously discussed, the choice of dataset, phrase count, and languages was based on a previous study by Sanchis-Trilles and Leiva (2014), who compared MemRep against other phrase sampling methods (NGRAMand random) and found that that MemRep was the best performer overall. In this experiment we compare Kaps against MemRep , for which we should use the same data of such previous study. Nonetheless, we believe that selecting movie subtitles for analysis is a good choice, as they reflect everyday language.

OpenSubtitles is a curated collection of parallel corpora extracted from the website http://opensubtitles.org. Being an open collection, anyone can contribute with their own subtitles at anytime. The OpenSubtitles dataset contains 2.8 million subtitle files in 60 languages for a total of over 17 billion tokens in 2.6 billion sentences, making it the world's largest multilingual corpus (Lison and Tiedemann, 2016). It is publicly available at http://opus.lingfil.uu.se/OpenSubtitles2016.php.

#### 4.5.2. Participants

We recruited 20 native speakers of each language, via specialized mailing lists, official language centers, language exchange groups, and italki.com (an online platform with teachers and native speakers in different languages). All participants had to sign a consent form were they had to verify that: they are at least 18 years old; they are native speakers of the language they choose to perform the study; they use a desktop PC or a laptop. Each participants was paid a 10 Euro gift voucher.

#### 4.5.3. Method

Participants were randomly presented with MemRep and Kaps phrases (they were not told which were which), ensuring that half of the phrases were derived from each sampling method. As in the previous small-scale user study, participants were shown a phrase for 5 s or until the first keystroke, whatever happened first. Afterward, the phrase disappeared and users had to write it (as much as they could remember) with their QWERTY keyboard. Phrases were lowercased and punctuations symbols were removed.

We tried to mitigate potential confounders of uncontrolled studies such as the one we conducted. For example, we ensured that users were not using a mobile device (via browser's user agent string), the phrases shown were not allowed to be copied and pasted (via JavaScript) and each user could only take the study once (via email and IP filtering). Eventually, each participant entered 100 phrases from both datasets, one phrase at a time, picked at random. This amounts to 20 participants × 100 phrases × 2 sampling techniques × 10 languages = 40,000 annotated phrases overall.

#### 4.5.4. Results

Figs. 7 and 8 summarize the results of this experiment. As can be observed, both sampling methods performed similarly, with memorization times around 2 seconds per phrase and character error rates ranging

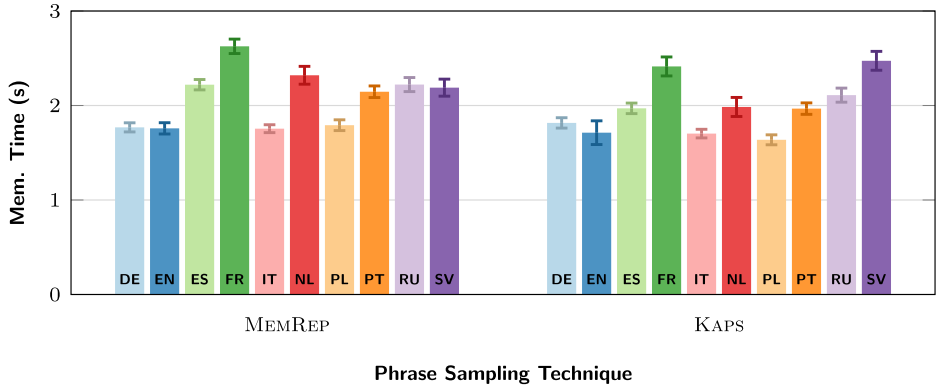


**Fig. 7.** Phrase memorization time. Error bars denote 95% CIs.

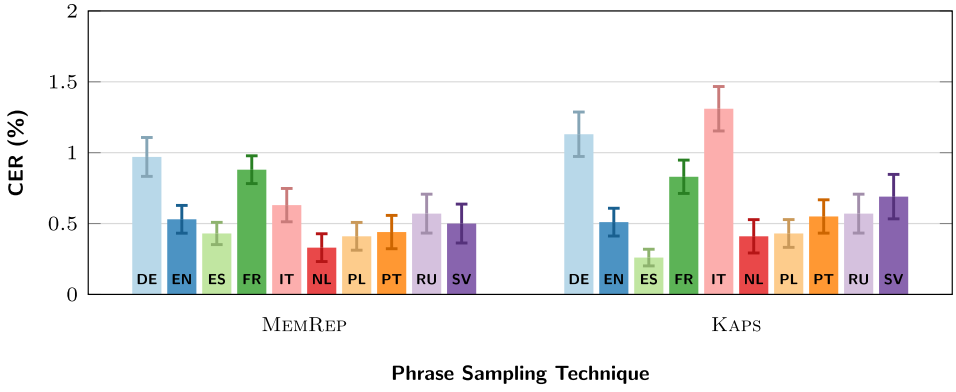


**Fig. 8.** Character error rate. Error bars denote 95% CIs.

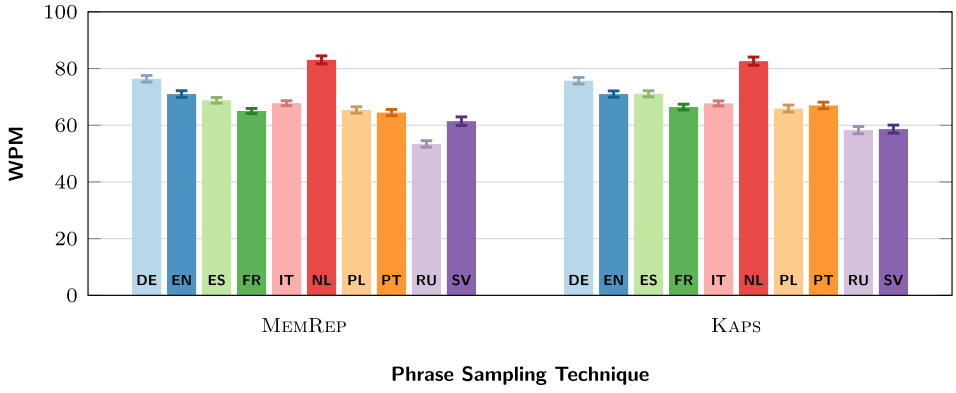


**Fig. 9.** Text entry speed, in words per minute. Error bars denote 95% CIs.

from 0.5% to ∼1%. If we compare the results with the Spanish language against the results from the language proficiency analysis (Section 4.1), we can see that both memorization times and CER values are in the same ranges.

We also report in Fig. 9 the text entry speeds, in words per minute, to provide an overview of the participants typing performance. For standardization purposes, a "word" is defined as "five consecutively entered characters, including spaces" (Yamada, 1908). As can be observed in the figure, participants achieved around 70 WPM, which suggests an average typing performance according to the text entry literature (Arif and Stuerzlinger, 2009).

We conducted a repeated measures two-way ANOVA (language and sampling method were factors) to assess the differences in our dependent variables of interest: memorization time and CER. The sampling method was a within-subject factor, whereas language was a between-subject factor; thus this is a mixed design. And since it is also a repeated measures design, we controlled for natural variation from participant to participant.

We observed a significant effect attributed to the language of the phrases: memorization time $\left[F_{(9,147)} = 2.91, p < .01, \eta_p^2 = 0.18\right]$; CER $\left[F_{(9,147)} = 11.28, p < .001, \eta_p^2 = 0.41\right]$. We also observed a significant effect attributed to the sampling method: memorization time $\left[F_{(1,146)} = 139.35, p < .001, \eta_p^2 = 0.49\right]$; CER $\left[F_{(1,146)} = 4.00, p < .05, \eta_p^2 = 0.03\right]$. In addition, a significant interaction between language and sampling method was found: memorization time $\left[F_{(9,146)} = 14.72, p < .01, \eta_p^2 = 0.48\right]$; CER $\left[F_{(9,146)} = 4.46, p < .001, \eta_p^2 = 0.22\right]$. Effect sizes suggest moderate practical importance.

We then split the data by language and analyzed the individual differences among each sampling method across languages. For this, we

used pairwise comparisons via *t*-tests (two-tailed, Bonferroni corrected). We observed that there were statistically significant differences in memorization time for MemRep phrases in French vs. English ($p < .05$). We also observed that there were statistically significant differences in CER for Kaps phrases in Spanish vs. Italian ($p < .01$). All other comparisons were not found to be significant.

We should point out that, contrary to the experiment conducted in Section 4.1, here all participants were *native* speakers of the language they evaluated. Thus, it was somewhat expected that they would tolerate semantic inaccuracies, something which Kaps does take into consideration but MemRep does not. In addition, we have observed that MemRep produces very consistent phrases overall, perhaps too much consistent, as pointed out by Leiva and Sanchis-Trilles (2014). For example, all MemRep phrases have exactly the same length for each language. This is actually an important observation, since phrase length is the most relevant feature to explain memorization time and CER: the longer the phrase the harder it is to memorize and the chance of typing mistakes increases with longer phrases (Vertanen and Kristensson, 2011b). On the contrary, Kaps produces phrases of variable length. In general, this behavior is consistent across features and languages, suggesting thus that Kaps phrases have more potential and are more flexible than those produced by MemRep .

Finally, it should be noted that Kaps is much simpler to use in practice than MemRep and allows text entry researchers to standardize the phrase sampling procedure by employing a single and common resource that is available in literally hundreds of languages.

## 5. General discussion and limitations

At a first glance, phrase sampling might seem like a side-track in text entry research. Is the choice of phrase sets the most urgent issue in text entry? Probably the overall focus of the field should be on improving text entry methods themselves. However, we have shown that language plays a crucial role on text entry performance. If users are not shown phrases in their native language, text entry metrics will be artificially inflated.

So far, phrases uses as stimuli in current text entry evaluations mainly exist in English and thus the evaluation of text entry methods in other languages is limited. Therefore, it is important that a sampling method like Kaps is language-agnostic, to enable truly multilingual text entry experiments. Further, because our sampling method is automated, it is expected to behave similarly in different languages.

The inability to evaluate text entry techniques in the native language of the users has long been considered a bottleneck in the Information and Communication Technologies for Development (ICT4D). Multilingual text entry research often lacks the rigor necessary to fully understand or evaluate novel text entry techniques due to the absence of adequate phrase sets. We hope that our experiments have convinced the reader that phrase sampling is an important problem in text entry, and especially critical to those evaluations involving non-English speakers.

### 5.1. Differences between Kaps and MemRep

Kaps was primarily motivated by the MemRep technique (Leiva and Sanchis-Trilles, 2014; Sanchis-Trilles and Leiva, 2014), therefore we believe it is worth discussing how Kaps builds upon and improves MemRep . In essence, the MemRep technique (Leiva and Sanchis-Trilles, 2014) also approaches phrase sampling as a ranking task based on CER prediction and, being model-driven, it is also language-agnostic. However, everything else is different in Kaps.

First, Kaps uses a completely different ranking approach, based on knowledge graphs, which allows to easily handle an abstract representation of words across languages. Second, Kaps uses a completely different set of model features, see Section 3.3 and Appendix B, which are much simpler to compute and therefore may have better performance in practice. Third, Kaps does not require an additional large dataset to learn a language model, just the input dataset from which phrases will be sampled, which simplifies the sampling procedure to a great extent. Finally, Kaps incorporates phrase complexity as a new dimension that agglutinates syntactics and semantics. This dimension is important because, as we have discussed in previous sections, ill-formed phrases, fragmented phrases, or phrases with no meaning may be confusing to participants.

### 5.2. Addressing potential treats to internal validity

This work enables us to better understand how language affect phrase sampling. Overall, the phrase language has been a largely neglected variable in text entry experiments. Nevertheless, while we have argued that language *is* an important variable to control for, there are other variables that might harm the validity of the experiment, such as age, users' individual skill, time of day, general tiredness, experimental setup, type of device, etc. Some of these effects could be mitigated e.g. by recruiting a larger or more diverse pool of participants, otherwise we should be aware that these variables could interact and introduce noise. In general, a good text entry experiment that cares about internal and external validity should be longitudinal and use a large number of phrases. Indeed, when the whole experiment consists of the entry of 10 or even fewer phrases per participant, the content of those phrases becomes a particularly important issue since it can have a significant effect on the results.

### 5.3. Balancing language performance

In multilingual text entry evaluations, where the experimenter is analyzing the performance of some text entry technique with speakers of different languages, it is important that the phrase sets have similar performance across languages. In this context the balance of words per phrase and letters per word are important measures to preserve. Otherwise, a language that has a higher ratio by default (cf. German vs.English) would result in lower text entry input speed based metrics, such as the usual words per minute, and possibly will increase the chance of committing more transcription errors. As shown in Table 10, Kaps manages to make these measures quite similar across languages: about 4 words per phrase and 4 letters per word on average. This suggests once more that our phrase sampling method would perform similarly in different languages.

Nevertheless, we should point out that these ratios will obviously not hold across *all* languages. They just appear to be feasible and desirable for our set of evaluated 10 languages, for the reasons stated above.

### 5.4. Does one model fit many languages?

Kaps uses a model learned from an English dataset to predict memorability (among other phrase features) in other languages. While we have shown that it works well with the ten languages under study, it seems natural to think that better models could be estimated using datasets in each specific language. Nevertheless, Kaps was designed to free text entry researchers from having to build their own models whenever they use a different language or text genre. Or is it worth the effort?

To address this concern, we performed a correlation analysis using the CER-labeled phrases resulting from the crowdsourced study with native speakers (see Section 4.5). We measured the Spearman correlation[15] between CER′ (predicted by Kaps) and the observed CER (committed by the participants). As shown in Table 5, the learned model correlates similarly with each of the ten languages, including English (which used a completely different phrase set). We should note that the correlation is weak because our native participants committed very few transcription errors (see Fig. 8) and Kaps cannot predict CER′ = 0% for any phrase

[15] The Spearman correlation is a nonparametric measure of rank correlation between two variables.

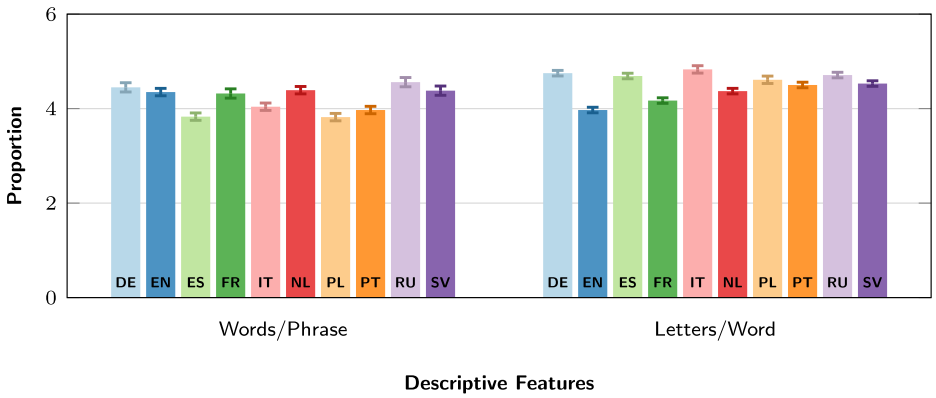


**Fig. 10.** Descriptive features of the phrases selected by KAPS for all the languages under study. Each language comprises 2,000 phrases. Error bars denote 95% confidence intervals.

**Table 5**
Correlation between the CER predicted by KAPS (CER′) and the observed CER over the phrases annotated in the crowdsourced study by native speakers.

| Language | DE | EN | ES | FR | IT | NL | PL | PT | RU | SV |
|---|---|---|---|---|---|---|---|---|---|---|
| Spearman's $\rho$ | 0.06 | 0.07 | 0.20 | 0.10 | 0.04 | 0.34 | 0.07 | 0.21 | 0.04 | 0.22 |

of non-zero length. We also should note that Spearman's $\rho$ is a measure of a monotonic relationship and thus it does not imply there is no relationship between CER′ and the observed CER. The message that we would like to convey to the reader, however, is that the learned model is versatile enough to be applied across the languages we have studied.

### 5.5. Working with more complex languages

We have shown that KAPS works well for a set of Indo-European languages, which represent an important portion of the current living languages (Lewis, 2013). However, KAPS is based on the statistical analysis of particular features, the importance of which may vary on a per-language basis. Therefore, KAPS may not be applicable to every possible language. For example, Chinese words are not even formed by individual letters and Hindi mostly requires two or more letters to be combined together to form a character.

Nevertheless, in most of the languages we have analyzed, a character may comprise diacritical marks like the umlaut (ä), acute accent (á), or cedilla (ç), something which KAPS can successfully account for. Therefore, to ensure that our model can work with more "complex" languages, the features should be either revisited or adapted. One plausible research direction in this regard is experimenting with labeled data in other languages or even use other memorability predictors, which could possibly lead to more general models and eventually would improve our sampling method. Looking forward, we hope that this work will inspire researchers and be suitable for use in a variety of text entry evaluations.

## 6. Conclusion and future work

This work represents a significant effort to *standardize* the evaluation mechanisms for text entry experiments across languages. We propose KAPS , a mathematically principled method for sampling text corpora, in which the goal is to select a smaller phrase set that seeks a balance among memorability, representativeness, and complexity; all of these are desired properties for conducting text entry evaluations. Under the hood, KAPS uses the BabelNet multilingual semantic network as a common resource.

This work also represents a significant effort to *simplify* the sampling procedure for conducting text entry experiments across languages. Contrary to its closest peer MEMREP , KAPS does not need an additional (large) text corpus as input, just the dataset from which phrases will be drawn. This facilitates the experimenter's work to a great extent, since she just has to decide how many phrases are needed for her text entry experiment.

Our empirical evidence suggests that text entry researchers should abandon their trusted phrase sets when conducting experiments in languages different from English, and switch to new ones in the native language of their participants. We have shown that KAPS behaves similarly across ten different languages and that results are comparable across them. Also, because KAPS is an automated sampling method, it can generalize to different tasks or domains. However, as discussed in the previous section, it must be noted that our features may not be applicable to every possible language.

Because KAPS is keyboard-agnostic, it can be used in any text entry evaluation. But evaluations focused e.g. on novel soft keyboards could include more design dimensions, such as word clarity (Yi et al., 2017) or the keyboard layout. For example, in some interfaces, user effort may be related to the costs of transitioning between keys (Paek and Hsu, 2011; Vertanen and Kristensson, 2011b). In such cases we might prefer a phrase subset in which combinations of characters appear about as often as in the target domain. To accomplish this, one could reorder or filter any of our phrase sets by following the procedure devised by Paek and Hsu (2011). Possible extensions to other keyboard layouts and text entry methods that do not rely on traditional keyboards could still be explored further. Another opportunity for future work consists in reviewing our phrase sets. For example, we have not looked for any offensive word in the phrases. As such, we recommend that researchers manually remove any undesirable phrases at their own discretion.

In general, our findings are of special relevance to text entry researchers interested in conducting experiments tailored to the linguistic capabilities of their participants, either in terms of the language itself or a particular text domain. We make our software and phrase sets (of 2,000 phrases each) in ten different languages publicly available at https://luis.leiva.name/kaps/. We expect researchers will use these phrase sets or will generate their own in their multilingual text entry research evaluations.

## Acknowledgments

We are very grateful to the UPV Language Center Office, the Valencia and Barcelona Language Exchange groups, the Corpora mailing list, and italki for helping us to reach native speakers of different languages. We thank Germán Sanchis-Trilles for fruitful discussions and Per Ola Kristensson and Keith Vertanen for reviewing a previous version of this manuscript. We also thank the anonymous IJHCS reviewers for their constructive feedback to strengthen the final version of this manuscript. This work assimilates some opinions and discussions we have had or overheard between SIGCHI colleagues in the past years. We thank them collectively.

## Appendix A. Examples of sampled phrases

Table A.6 provides an overview of the phrases sampled by KAPS from the OPENSUBTITLES dataset. We have made these phrase sets (2,000 phrases for each language) publicly available in two forms: full phrases (as in the original dataset) and tokenized + lowercased + no punctuation (as shown in the table).

**Table A.6**
KAPS phrase examples to conduct text entry experiments in different languages. The 'Rank' column indicates examples from the top (+) and bottom (−) positions of the ranking.

| Language | Rank | Sample phrases |
|---|---|---|
| German | + | es gibt kein ende |
| | + | komm und hilf mir |
| | - | können wir gehen endlich |
| | - | ich bring was zu essen |
| English | + | and i write novels |
| | + | a great city |
| | - | yes i mean no |
| | - | to cheer and comfort me |
| Spanish | + | quiero volver a tocar |
| | + | si le despierta volverá |
| | - | sí odio cuando pasa eso |
| | - | o lo que le hacen prometer |
| French | + | fais la tourner |
| | + | c'était une petite ville |
| | - | ça va être demain |
| | - | allô je peux parler à lisa |
| Italian | + | gia che fai |
| | + | permettimi di farlo per te |
| | - | come te ia passi |
| | - | chi offre di più |
| Dutch | + | bescherm hem met je leven |
| | + | kom en zie |
| | - | mag ik larry zeggen |
| | - | te vallen zoals een bom |
| Polish | + | jestem w mieście |
| | + | nie ma sensu starać się to zmienić |
| | - | nic wam nie będzie |
| | - | nie mogę pani powiedzieć |
| Portuguese | + | não tens como mudar |
| | + | pode se dizer |
| | - | deixa que olhem |
| | - | isso pode funcionar |
| Russian | + | знаю что ты сделал |
| | + | она была в городе пару дней назад |
| | - | дай мне свой телефон |
| | - | он сказал я цитирую |
| Swedish | + | gå bara ur |
| | + | jag ställer in timern |
| | - | nej för att vi slösar bort tid |
| | - | era ord är ren poesi |

## Appendix B. List of knowledge graph-based candidate features

Table B.7 shows all the features that were considered before performing feature selection. The 'Select' column indicates the final set of features, denoted with the acronym we used in Section 3.3.

**Table B.7**
List of the 78 features considered, before performing feature selection, marking in bold typeface those that were selected by the final model.

| Feature name | Select |
|---|---|
| NUM_WORDS | **Nw** |
| AVG_NUM_CHARS_PER_WORD | **Achr** |
| STDEV_NUM_CHARS_PER_WORD | |
| RATIO_WORDS_WITH_NUMBERS | **Pn** |
| NUM_SOURCE_WORDS | |
| NUM_SOURCE_WORDS_WITH_SYNSETS | |
| RATIO_SOURCE_WORDS | |
| RATIO_SOURCE_WORDS_WITH_SYNSETS | |
| RATIO_WORDS_WITH_SYNSETS | **PwWs** |
| SOURCE_SYNSETS | |
| SOURCE_UNIQUE_SYNSETS | |
| AVG_SOURCE_SYNSETS_PER_WORD | |
| MEDIAN_SOURCE_SYNSETS_PER_WORD | |
| STDEV_SOURCE_SYNSETS_PER_WORD | |
| MIN_SOURCE_SYNSETS_PER_WORD | |
| MAX_SOURCE_SYNSETS_PER_WORD | |
| NUM_PATHS | |
| AVG_PATHS_PER_SOURCE_WORD | |
| AVG_PATHS_PER_SOURCE_SYNSET | |
| MEDIAN_PATHS_PER_SOURCE_WORD | |
| MEDIAN_PATHS_PER_SOURCE_SYNSET | |
| STDEV_PATHS_PER_SOURCE_WORD | |
| STDEV_PATHS_PER_SOURCE_SYNSET | |
| MIN_PATHS_PER_SOURCE_SYNSET | |
| MAX_PATHS_PER_SOURCE_SYNSETT RATIO_PATHS_LEN_0 | |
| RATIO_PATHS_LEN_1 | |
| RATIO_PATHS_LEN_2 | |
| RATIO_PATHS_LEN_3 | |
| NUM_PATHS_LEN_0 | |
| NUM_PATHS_LEN_1 | |
| NUM_PATHS_LEN_2 | |
| NUM_PATHS_LEN_3 | |
| NUM_EXPANDED_SYNSETS | |
| NUM_NODES | |
| RATIO_EXPANDED_SYNSETS | |
| NUM_EXPANDED_SYNSETS | |
| NUM_EDGES | |
| AVG_INOUT_DEGREE | **Aio** |
| MEDIAN_INDEGREE | |
| MEDIAN_OUTDEGREE | |
| STDEV_OUTDEGREE | **SDo** |
| STDEV_INDEGREE | **SDi** |
| MIN_OUTDEGREE | |
| MAX_OUTDEGREE | **Mo** |
| MIN_INDEGREE | |
| MAX_INDEGREE | **Mi** |
| RATIO_NODES_EDGES | |
| SUM_EDGE_WEIGHT | |
| AVG_EDGE_WEIGHT | |
| MEDIAN_EDGE_WEIGHT | |
| STDEV_EDGE_WEIGHT | |
| MIN_EDGE_WEIGHT | |
| MAX_EDGE_WEIGHT | |
| RATIO_WORD_VERBS | |
| | **PwWvS** |
| RATIO_WORD_NOUNS | |
| RATIO_WORD_ADVERBS | |
| RATIO_WORD_ADJECTIVES | |
| RATIO_SOURCE_SYNSET_VERBS | |
| RATIO_SOURCE_SYNSET_NOUNS | |
| RATIO_SOURCE_SYNSET_ADVERBS | |
| RATIO_SOURCE_SYNSET_ADJECTIVES | |
| RATIO_GRAPH_SYNSET_VERBS | |
| RATIO_GRAPH_SYNSET_NOUNS | |
| RATIO_GRAPH_SYNSET_ADVERBS | |
| RATIO_GRAPH_SYNSET_ADJECTIVES | |
| NUM_WORD_VERBS | |
| NUM_WORD_NOUNS | |
| NUM_WORD_ADVERBS | |
| NUM_WORD_ADJECTIVES | |
| NUM_SOURCE_SYNSET_VERBS | |
| NUM_SOURCE_SYNSET_NOUNS | |
| NUM_SOURCE_SYNSET_ADVERBS | |
| NUM_SOURCE_SYNSET_ADJECTIVES | |
| NUM_GRAPH_SYNSET_VERBS | |
| NUM_GRAPH_SYNSET_NOUNS | |
| NUM_GRAPH_SYNSET_ADVERBS | |
| NUM_GRAPH_SYNSET_ADJECTIVES | |

## References


Arif, A.S., Stuerzlinger, W., 2009. Analysis of text entry performance metrics. In: Proceedings of IEEE International Conference on Science and Technology for Humanity (TIC-STH), pp. 100–105.

Bates, E., Kintsch, W., Fletcher, C.R., Giuliani, V., 1980. The role of pronominalization and ellipsis in texts: Some memory experiments. J. Exp. Psychol.-Learn. Mem. Cogn 6 (6), 676–691.

Burnham, K.P., Anderson, D.R., 2004. Multimodel inference: understanding AIC and BIC in model selection. Socio. Meth. Res 33 (2), 261–304.

Chomsky, N., 1955. Logical Structure of Linguistic Theory, 1975th edition Springer.

Coleman, M., Liau, T.L., 1975. A computer readability formula designed for machine scoring. J. Appl. Psychol 60 (1), 283–284.

Danescu-Niculescu-Mizil, C., Cheng, J., Kleinberg, J., Lee, L., 2012. You had me at hello: How phrasing affects memorability. In: Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), pp. 892–901.

Dunlop, M.D., Masters, M.M., 2009. Pickup usability dominates: a brief history of mobile text entry research and adoption. Int. J. Mobile Human Comput. Interact. (IJMHCI) 1 (1), 42–59.

Ferret, O., 1998. How to thematically segment texts by using lexical cohesion? In: Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1481–1483.

Flesch, R., 1948. Marks of a readable style. Columbia University Ph.d. thesis.

Franco-Salvador, M., Cruz, F.L., Troyano, J.A., Rosso, P., 2015. Cross-domain polarity classification using a knowledge-enhanced meta-classifier. Knowl.-Based Syst 86, 46–56.

Franco-Salvador, M., Rosso, P., Montes y Gómez, M., 2016. A systematic study of knowledge graph analysis for cross-language plagiarism detection. Inform. Process. Manag 52 (4), 550–570.

Franco-Salvador, M., Rosso, P., Navigli, R., 2014. A knowledge-based representation for cross-language document retrieval and categorization. In: Proceedings of European Chapter of the Association for Computational Linguistics (EACL), pp. 414–423.

Ganslandt, S., Jörwall, J., Nugues, P., 2009. Predictive text entry using syntax and semantics. In: Proceedings of International Conference on Parsing Technology (IWPT), pp. 37–48.

Ghosh, S., Joshi, A., 2014. Text entry in Indian languages on mobile: user perspectives. In: Proceeding of India HCI (IHCI), pp. 55–63.

González, I.E., Wobbrock, J.O., Chau, D.H., Faulring, A., Myers, B.A., 2007. Eyes on the road, hands on the wheel: thumb-based interaction techniques for input on steering wheels. In: Proceeding of Graphics Interface (GI), pp. 95–102.

Gunning, R., 1952. The Technique of Clear Writing. McGraw-Hill.

Heck, L.P., Hakkani-Tür, D., Tür, G., 2013. Leveraging knowledge graphs for web-scale unsupervised semantic parsing. In: Proceedings of INTERSPEECH, pp. 1594–1598.

Isokoski, P., Linden, T., 2004. Effect of foreign language on text transcription performance: finns writing English. In: Proceedings of Nordic Conf. on Human-Computer Interaction (NordiHCI), pp. 109–112.

Isokoski, P., Raisamo, R., 2000. Device independent text input: a rationale and an example. In: Proceedings of Working Conference on Advanced Visual Interfaces (AVI), pp. 76–83.

James, P., 1992. Knowledge graphs. In: Linguistic Instruments in Knowledge Engineering, pp. 97–117.

Just, M.A., Carpenter, P.A., 1980. A theory of reading: from eye fixations to comprehension. Psychol. Rev 87 (1), 329–354.

Kano, A., Read, J.C., Dix, A., 2006. Children's phrase set for text input method evaluations. In: Proceedings of Nordic Conference on Human-Computer Interaction (NordiHCI), pp. 449–452.
Kanungo, T., Orr, D., 2009. Predicting the readability of short web summaries. In: Proceedings of International Conference on Web Search and Data Mining (WSDM), pp. 202–211.
Karat, C.-M., Halverson, C., Horn, D., Karat, J., 1999. Patterns of entry and correction in large vocabulary continuous speech recognition systems. In: Proceedings of SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 568–575.
Keller, F., 2004. The entropy rate principle as a predictor of processing effort: an evaluation against eye-tracking data. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 317–324.
Klare, G.R., 1976. A second look at the validity of readability formulas. J. Reading Behav 8, 129–152.
Kristensson, P.O., Vertanen, K., 2012. Performance comparisons of phrase sets and presentation styles for text entry evaluations. In: Proceedings of International Conference on Intelligent User Interfaces (IUI), pp. 29–32.
Kruskal, W., Mosteller, F., 1979. Representative sampling, II: scientific literature, excluding statistics. In: International Statistical Review, Vol. 47, pp. 111–127.
Költringer, T., Isokoski, P., Grechenig, T., 2007. TwoStick: writing with a game controller. In: Proceedings of Graphics Interface (GI), pp. 103–110.
Leiva, L.A., Sanchis-Trilles, G., 2014. Representatively memorable: sampling the right phrase set to get the text entry experiment right. In: Proceedings of SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 1709–1712.
Lewis, M.P. (Ed.), 2013. Ethnologue: Languages of the World, Seventeenth. SIL International.
Lison, P., Tiedemann, J., 2016. OpenSubtitles2016: extracting large parallel corpora from movie and TV subtitles. In: Proceedings of Language Resources and Evaluation Conference (LREC), pp. 923–929.
Lyons, K., Clawson, J., 2012. Analytical evaluation of the impact of phrase set on text entry rates. Proceeding of Extended Abstracts on Human Factors in Computing Systems (CHI EA). Workshop on Designing ad Evaluating Text Entry Methods
Mackenzie, I.S., Felzer, T., 2010. SAK: scanning ambiguous keyboard for efficient one-key text entry. ACM T. Hum.-Comput. Interact. (TOCHI) 17 (3), 11:1–11:39.
MacKenzie, I.S., Soukoreff, R.W., 2002. Text entry for mobile computing: models and methods, theory and practice. Hum.-Comput. Interact 17 (2), 147–198.
MacKenzie, I.S., Soukoreff, R.W., 2003. Phrase sets for evaluating text entry techniques. In: Proceeding of Extended Abstracts on Human Factors in Computing Systems (CHI EA), pp. 754–755.
McLaughlin, G.H., 1969. SMOG grading: a new readability formula. J. Reading 12, 639–646.
Mihalcea, R., Radev, D., 2011. Graph-based natural language processing and information retrieval. Cambridge University Press.
Misra, H., Cappé, O., Yvon, F., 2008. Using lda to detect semantically incoherent documents. In: Proceeding of Conference on Computational Natural Language Learning (CoNLL), pp. 41–48.
Moro, A., Raganato, A., Navigli, R., 2014. Entity linking meets word sense disambiguation: a unified approach. Trans. Assoc. Comput. Linguist. (TACL) 2, 231–244.
Mostafazadeh, N., Mirroshandel, S.A., Ghassem-Sani, G., Babarsad, O.B., 2011. Pazesh: a graph-based approach to increase readability of automatic text summaries. In: Proceeding of Canadian Conference on Advances in Artificial Intelligence (Canadian AI), pp. 313–318.
Naseem, T., Barzilay, R., 2011. Using semantic cues to learn syntax. In: Proceeding of AAAI Conference on Artificial Intelligence (AAAI), pp. 902–907.
Navigli, R., Lapata, M., 2010. An experimental study of graph connectivity for unsupervised word sense disambiguation. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 32 (4), 678–692.
Navigli, R., Ponzetto, S.P., 2012. BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. 193, 217–250.
Navigli, R., Ponzetto, S.P., 2012. BabelRelate! a joint multilingual approach to computing semantic relatedness. In: Proceeding of AAAI Conference on Artificial Intelligence (AAAI), pp. 108–114.
Nelder, J.A., Weddrburn, R.W.M., 1972. Generalized linear models. J. R. Statist. Soc 135 (3), 370–384.
Paek, T., Hsu, B.-J.P., 2011. Sampling representative phrase sets for text entry experiments: a procedure and public resource. In: Proceeding of SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 2477–2480.
Page, L., Brin, S., Motwani, R., Winograd, T., 1998. The PageRank citation ranking: bringing order to the web. Technical report. Stanford Digital Library Technologies Project.
Popping, R., 2003. Knowledge graphs and network text analysis. Soc. Sci. Inf 42 (1), 91–106.
Prince, V., Labadié, A., 2007. Text segmentation based on document understanding for information retrieval. In: Proceeding of International Conference on Applications of Natural Language to Information Systems (NLDB), pp. 295–304.
Radev, D.R., Jing, H., Budzikowska, M., 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In: Proceeding of NAACL-ANLP-AutoSum – Workshop on Automatic Summarization, pp. 21–30.
Rayner, K., Duffy, S.A., 1986. Lexical complexity and fixation times in reading: effects of word frequency, verb complexity, and lexical ambiguity. Mem. Cognit 14 (3), 191–201.
Rubin, D.C., 1977. Very long-term memory for prose and verse. J. Verbal Learning Verbal Behav 16 (5), 611–621.
Sanchis-Trilles, G., Leiva, L.A., 2014. A systematic comparison of 3 phrase sampling methods for text entry experiments in 10 languages. In: Proceeding of ACM Conference on Human-computer interaction with mobile devices and services (MobileHCI), pp. 537–542.
Schwarz, G.E., 1978. Estimating the dimension of a model. Ann. Stat 6 (2), 461–464.
Senter, R.J., Smith, E.A., 1967. Automated readability index. Tech. Rep. AMRL-TR-6620. Wright-Patterson Air Force Base
Stocky, T., Faaborg, A., Lieberman, H., 2004. A commonsense approach to predictive text entry. In: Proceeding of SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 24–29.
Sweller, J., 1994. Cognitive load theory, learning difficulty, and instructional design. Learn. Instr 4 (4), 295–312.
Tanaka-Ishii, K., Hayakawa, D., Takeichi, M., 2003. Acquiring vocabulary for predictive text entry through dynamic reuse of a small user corpus. In: Proceeding of Annual Meeting of the Association for Computational Linguistics (ACL), pp. 407–414.
Thorn, A., Page, M., 2009. Interactions Between Short-Term and Long-Term Memory in the Verbal Domain. Psychology Press.
Vasiljevas, M., Šalkevičius, J., Gedminas, T., Damaševičius, R., 2015. A prototype gaze–controlled speller for text entry. In: Proceeding of International Symposium for Young Scientists in Technology, Engineering and Mathematics (SYSTEM), pp. 79–83.
Vertanen, K., Kristensson, P.O., 2011. The imagination of crowds: conversational AAC language modeling using crowdsourcing and large data sources. In: Proceeding of Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 700–711.
Vertanen, K., Kristensson, P.O., 2011. A versatile dataset for text entry evaluations based on genuine mobile emails. In: Proceeding of ACM Conference on Human-computer interaction with mobile devices and services (MobileHCI), pp. 295–298.
Vertanen, K., Kristensson, P.O., 2014. Complementing text entry evaluations with a composition task. ACM T. Hum.-Comput. Interact. (TOCHI) 21 (2), 8:1–8:33.
Wobbrock, J.O., Myers, B.A., 2006. Analyzing the input stream for character-level errors in unconstrained text entry evaluations. ACM T. Hum.-Comput. Interact. (TOCHI) 13 (4), 458–489.
Yamada, H., 1908. A historical study of typewriters and typing methods: from the position of planning Japanese parallels. J. Information Processing (JIPS) 2 (4), 175–202.
Yi, X., Yu, C., Shi, W., Bi, X., Shi, Y., 2017. Word clarity as a metric in sampling keyboard test sets. In: Proceeding of SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 4216–4228.
Yoshimi, T., Okunishi, T., Yamaji, T., Fukumochi, Y., 1998. Evaluation of importance of sentences based on connectivity to title. In: Proceeding of International Conference on Computational Linguistics (COLING), pp. 1443–1447.
Zajic, D.M., Lin, J., Dorr, B.J., Schwartz, R., 2006. Sentence compression as a component of a multi-document summarization system. In: Proceeding of Document Understanding Conference (DUC) Workshop, pp. 1–7.
Zhai, S., Kristensson, P.O., Smith, B.A., 2005. In search of effective text input interfaces for off the desktop computing. Interact. Comput 17 (3), 229–250.
Zhai, S., Sue, A., Accot, J., 2002. Movement model, hits distribution and learning in virtual keyboarding. In: Proceeding of SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 17–24.
Zhu, L., Gao, S., Pan, S.J., Li, H., Deng, D., Shahabi, C., 2013. Graph-based informative-sentence selection for opinion summarization. In: Proceeding of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 408–412.