Creating vocabulary exercises through NLP

Manex Agirrezabal[1][0000−0001−5909−2745],
Begoña Altuna[2][0000−0002−4027−2014], Lara Gil-Vallejo[3][0000−0002−9393−5651], Josu Goikoetxea[2][0000−0001−5568−4014], and Itziar Gonzalez-Dios[2][0000−0003−1048−5403]

[1] University of Copenhagen,
Department of Nordic Studies and Linguistics. Manex.Aguirrezabal@hum.ku.dk
[2] Ixa Group, University of the Basque Country, Manuel Lardizabal
1, 20018 Donostia. {begona.altuna,josu.goikoetxea,itziar.gonzalezd}@ehu.eus
[3] Universitat Oberta de Catalunya
lgilva@uoc.edu

Abstract. The use of technologies in Humanities opens new research opportunities as it allows the access to vast amounts of data such as textual corpora. As, in the Digital Humanities domain, a considerable amount of the research is done on digitised corpora, Natural Language Processing tools can be of much help in their exploitation for they help extracting linguistic information. We present a series of experiments in which we propose text transformations to generate vocabulary learning exercises based on Natural Language Processing. We describe the corpus, databases and tools we have employed in our approach and we offer an overview of a multilingual language processing pipeline. Then we present the experiments and their output. We finally discuss the strengths and shortcomings of our approach.

Keywords: Natural Language Processing·Text transformation·Vocabulary learning.

1 Introduction

The increasing use of Information and Communication Technologies has opened new research possibilities and the growing amount of digital data and the development of processing tools have changed the paradigms of many research fields. In the case of Humanities, the so-called Digital Humanities (DH) aim at exploiting the vast amounts of digitised corpora with the help of the Natural Language Processing (NLP) tools among others. In fact, DH and NLP can be considered closely related fields as Humanities are often based on textual data and language knowledge. NLP comprehends a wide range of research interests and approaches that can be useful in DH in the NLP's aim of providing computers with human language knowledge.

Since the 60's, NLP's challenges have mainly been Speech Recognition, Natural Language Understanding and Natural Language Generation and some of its main purposes are spell checking, parsing, machine translation, information retrieval, question answering. However, it also addresses some more advanced applications such as assistance for human creativity and uses in education. Our work can be placed in those advanced uses of NLP.

In what concerns enhancing human creativity, the so-called computational creativity is a notably prominent field. Computational creativity aims at modelling, simulating or replicating human creativity using computational models and, hence, it can help to enrich human productiveness providing suggestions, as it is the case of automatic poetry generation [16]. For example, Agirrezabal et al. [4] describe word substitutions based on Part-of-Speech and meaning in order to create new poems form existing ones preserving the coherence of the texts. Although that approach is focused on poetry, the substitutions presented can be of much help in other tasks dealing with vocabulary substitution.

Regarding the educational usage, one should consider that NLP has also largely been used for exercise generation and for student assessment [5, 18]. In fact, automatising exercise generation can be considered a hybridisation of both computational creativity and educational use of NLP, as automatic exercise generation may offer suitable options where human creativity struggles —it can save time to teachers and textbooks designers or can be useful for self-learning.

Our experimentation is centred, thus, on reusing NLP tools for computational creativity and educational purposes. In this paper we present a proposal of a set of NLP resources and tools for text adaptations to be used in the areas of language teaching and inclusion. More precisely, we focus on text transformations that can be used in the language acquisition field e.g. creating vocabulary exercises.

We strongly aim at the reusability of our proposals and, hence, all the resources and tools used in our work (corpora, databases and tools) are freely available. Moreover, we provide the code we have created[4] under the *Creative Commons* licence, international version (4.0) and NonCommercial attribution (CC 4.0 BY-NC).

This paper is structured as follows: in Sections 2 and 3 we describe the resources and tools we use in our experiments; in Section 4 we present the texts transformations we propose and in Section 5 we discuss certain issues arisen from our experiments. Finally, we conclude and outline the future work in Section 6.

2 Linguistic Resources: Corpus and Databases

Our work is prominently based on text and, precisely, our two main resources are a narrative corpus collected by us (Section 2.1) and the WordNet [14] database (Section 2.2). In addition, we have also taken advantage of the ImageNet [13] database, in which images are related to concepts and mapped according to those (Section 2.3).

2.1 Fairy tales Corpus

In our experiment, we have opted for children's literature to build our corpus; more precisely fairy tales we have extracted from Project Gutenberg[5] and Wikipedia. Project Gutenberg offers over 57,000 freely available e-books in 67 languages. In the case of Wikipedia, each tale is indexed as an individual Wikipedia entry which offers

[4] https://github.com/dss2016eu/codefest/tree/master/nlp_lac

[5] http://www.gutenberg.org

the background of the story as well a version of the tale. The reasons for opting for well-known fairy tales are the following:

- Folktales have been widely employed in education [19].
- They are optimal for language learning for they are commonly told in easy language and are widely known [19].
- There is a wide range of fairy tales freely available.
- Versions of those tales can be easily found for a myriad of languages and multilingual approaches can be easily implemented.
- Text from Project Gutenberg and Wikipedia can be easily obtained as plain unformatted texts, which simplifies the textual preprocessing stage.

In order to give an idea of the size of the corpus we have compiled, we have listed the token amounts for each tale and language in Table 1. As can be noticed, we have tried to gather texts of similar length in order to achieve comparable results in different languages.

Language	**Little Red Riding Hood**	**Hansel and Gretel**	**Total**
English	1384	2870	4254
Spanish	564	2637	3201
Catalan	563	647	1210
Galician	497	93	590
French	1432	2680	4112
German	1257	2663	3920
Italian	1199	1946	3145
Portuguese	662	2610	3272
Dutch	1311	2726	4037
Basque	274	174	448

Table 1. Tokens per document and language

In this paper, we will illustrate our work through selected passages from the *Little Red Riding Hood* tale by the Grimm brothers.

2.2 WordNet

WordNet[6] is a large lexical database of English [14] where nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets). For example, the words *car, auto, automobile, machine,* and *motorcar* are grouped in a synset which denotes the concept *a motor vehicle with four wheels; usually propelled by an internal combustion engine*. Moreover, the synsets are related among them. The most important semantic relations are hypernymy-hyponymy, meronymy, troponymy (for verbs) and antonymy. In Figure 1 we present the synset *car, auto, automobile, machine, motorcar* and its related words.

[6] http://wordnetweb.princeton.edu/perl/webwn

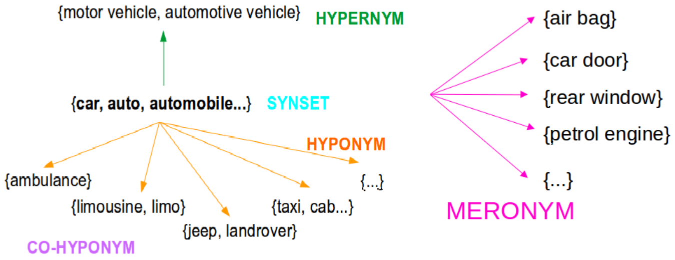


Fig. 1. Main semantic relations in WordNet

However, it should be taken into account that a word may have more than one meaning. When entering a query for a certain word we find all the word's senses listed. E.g. if we look for the noun *hood*, we may find these three senses among others: i) *a headdress that protects the head and face* ii) *protective covering consisting of a metal part that covers the engine* and iii) *the folding roof of a carriage*. Hence, it is plausible for a word to appear in more than one synset depending on which sense is considered.

Following the English WordNet philosophy, WordNets for many languages have been developed. For example, during the EuroWordNet (EWN) project[7] WordNets of several European languages (English, Dutch, Italian, Spanish, German, French, Czech and Estonian) were created. Nevertheless, these WordNets for different languages are not isolated databases. The Inter-Lingual-Index (ILI) was created to provide an efficient mapping across the autonomous WordNets. Via this index, languages are interconnected so that it is possible to go from the words in one language to the equivalent words in any other language listed.

The list of available WordNets for different languages has been increasing since then. For example, the Open Multilingual WordNet [9] (OMW)[8] —a product of the Global WordNet Association[9]— provides access to open WordNets in over 150 languages, all linked to the English WordNet.

In this work, we have chosen WordNet for our experimentation because i) it is a well-known multilingual and ii) a freely available resource and iii) it has been used in many NLP applications. Specifically, we have used the OMW that is included in NLTK [8], which is a suite of Python libraries and software for symbolic and statistical NLP, and we have used it to substitute words (the lemmas, exactly) with semantically related concepts or with its equivalents in other languages.

2.3 ImageNet

ImageNet[10] [13] is a large-scale image database arranged according to the hierarchy in WordNet. It contains images for nouns and each node of the hierarchy is represented by thousands of images. That is to say, nouns in the English WordNet get images

[7] http://projects.illc.uva.nl/EuroWordNet/
[8] http://compling.hss.ntu.edu.sg/omw/
[9] http://globalwordnet.org/
[10] http://www.image-net.org

that represent them. More precisely, the ImageNet project aims at offering 500-1000 images for each synset. As WordNet, ImageNet is freely available and ready to use.

Although ImageNet was initially developed for visual information processing and tasks such as non-parametric object recognition, tree-based image classification and automatic object localization, it has been proven very useful in our experiment. As a matter of fact, in this work we have employed ImageNet for the substitution of nouns in texts for images. By means of that, we have been able to create texts with images similar to texts with pictograms. We further describe that approach in Section 4.2.

3 Preprocessing with NLP tools

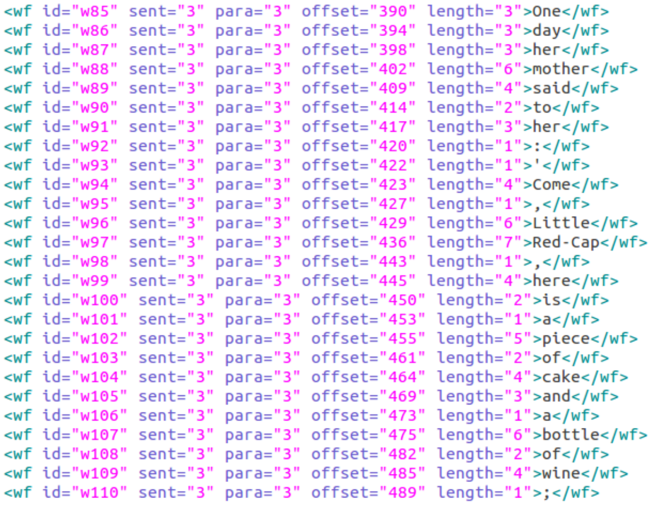
Textual processing has been done by some existing off-the-shelf NLP tools. The required processing for most of the languages (Basque, Dutch, English, French, German, Italian and Spanish) has been done through Ixa-pipes[11] [1], whereas the processing for the rest of languages (Galician, Catalan and Portuguese) has been conducted through FreeLing[12] [11]. Although we have employed two different processing pipelines, the steps are comparable. As a consequence, we will describe the processing modules relevant to this work based on Ixa-pipes' performance.

Ixa-pipes is a modular chain of NLP tools (or pipes) which provide easy access to NLP technology for several languages. Modular means that the different processing modules (for specific linguistic analysis tasks) can be chosen according to the needs of each experiment and that new modules can be added to address new needs. We present the processes we have carried out below:

- **Tokenisation:** it is the process of splitting sequences of characters into minimal meaningful units. In the tokenisation process, texts are divided into words, numbers, acronyms or punctuation marks. As can be seen in Figure 2, the sentence (*sent="3"*) has been split into tokens (*wf*) and each token has been assigned an identifier (*id*). Tokenisation parameters are defined for each language so as to take into account the special characters and singular orthography and punctuation rules each language may have. It is also to be pointed out that punctuation and hyphenation exceptions have been taken into account as can be seen for the *Red-Cap* token, which has been considered a single unit despite the fact there is a hyphen involved.
- **Lemmatisation:** it consists in removing word inflection to return the dictionary form or *lemma* of a word. For example, from the verb form *is* we obtain the lemma *be* after lemmatisation is done. In Ixa-pipes lemmatisation is performed by lexical look-up methods in which each word in text is checked in a dictionary. The lemma is the basis in our experimentation, since we create the transformations based on it. Consequently, an unknown or incorrect lemma can lead to errors in the following processes.
- **Part-of-Speech tagging:** it consists in assigning a grammatical category to each of the tokens. In Ixa-pipes this is a two step procedure: first, all the possible

[11] http://ixa2.si.ehu.es/ixa-pipes

[12] http://nlp.lsi.upc.edu/freeling/node/1



```
<wf id="w85" sent="3" para="3" offset="390" length="3">One</wf>
<wf id="w86" sent="3" para="3" offset="394" length="3">day</wf>
<wf id="w87" sent="3" para="3" offset="398" length="3">her</wf>
<wf id="w88" sent="3" para="3" offset="402" length="6">mother</wf>
<wf id="w89" sent="3" para="3" offset="409" length="4">said</wf>
<wf id="w90" sent="3" para="3" offset="414" length="2">to</wf>
<wf id="w91" sent="3" para="3" offset="417" length="3">her</wf>
<wf id="w92" sent="3" para="3" offset="420" length="1">:</wf>
<wf id="w93" sent="3" para="3" offset="422" length="1">'</wf>
<wf id="w94" sent="3" para="3" offset="423" length="4">Come</wf>
<wf id="w95" sent="3" para="3" offset="427" length="1">,</wf>
<wf id="w96" sent="3" para="3" offset="429" length="6">Little</wf>
<wf id="w97" sent="3" para="3" offset="436" length="7">Red-Cap</wf>
<wf id="w98" sent="3" para="3" offset="443" length="1">,</wf>
<wf id="w99" sent="3" para="3" offset="445" length="4">here</wf>
<wf id="w100" sent="3" para="3" offset="450" length="2">is</wf>
<wf id="w101" sent="3" para="3" offset="453" length="1">a</wf>
<wf id="w102" sent="3" para="3" offset="455" length="5">piece</wf>
<wf id="w103" sent="3" para="3" offset="461" length="2">of</wf>
<wf id="w104" sent="3" para="3" offset="464" length="4">cake</wf>
<wf id="w105" sent="3" para="3" offset="469" length="3">and</wf>
<wf id="w106" sent="3" para="3" offset="473" length="1">a</wf>
<wf id="w107" sent="3" para="3" offset="475" length="6">bottle</wf>
<wf id="w108" sent="3" para="3" offset="482" length="2">of</wf>
<wf id="w109" sent="3" para="3" offset="485" length="4">wine</wf>
<wf id="w110" sent="3" para="3" offset="489" length="1">;</wf>
```

Fig. 2. Tokenisation example for *Little Red Riding Hood*

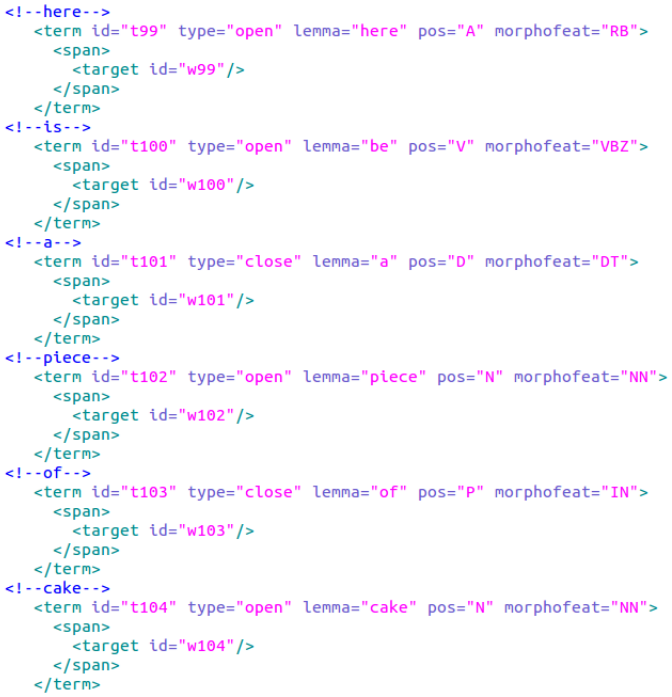
analysis are assigned to each token and then, the most suitable one is selected. In this process both linguistic knowledge (rules) and statistical methods are combined. In Figure 3, we present the complete annotation of a segment of the *Little Red Riding Hood* tale in English. One may notice that each token is presented in blue, lemmas are expressed by the *lemma* attribute and Part-of-Speech (PoS) information is given in the *pos* attribute.

- **Word Sense Disambiguation (WSD)**[2]**:** it is a NLP task that aims to identify the sense of a word in a sentence when that word has more than one sense. For example, given the sentence *We took off the hood*, the goal is to assess whether the word *hood* refers to the headdress or to the car cover. In order to perform WSD in this work, we have used the state-of-the-art tool UKB [3], which is also integrated in Ixa-pipes, and works with English, Basque, Bulgarian, Portuguese and Spanish. As an output, UKB offers all the possible WordNet synsets (*reference*) with a confidence value (*confidence* attribute) as we show in Figure 4 for the word *hood*.

This is the preprocessing needed in order to carry out the text transformations presented in Section 4.

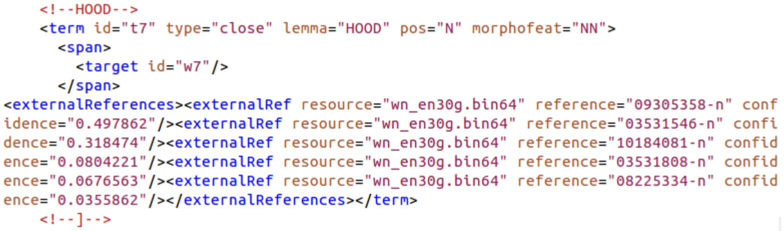
4 Text Adaptations

Once textual processing has been done, we have profited from the extracted linguistic information to alter texts and generate reading and vocabulary activities automatically. In the following subsections we describe the kind of exercises we have created for helping to acquire language.



```
<!--here-->
   <term id="t99" type="open" lemma="here" pos="A" morphofeat="RB">
     <span>
       <target id="w99"/>
     </span>
   </term>
<!--is-->
   <term id="t100" type="open" lemma="be" pos="V" morphofeat="VBZ">
     <span>
       <target id="w100"/>
     </span>
   </term>
<!--a-->
   <term id="t101" type="close" lemma="a" pos="D" morphofeat="DT">
     <span>
       <target id="w101"/>
     </span>
   </term>
<!--piece-->
   <term id="t102" type="open" lemma="piece" pos="N" morphofeat="NN">
     <span>
       <target id="w102"/>
     </span>
   </term>
<!--of-->
   <term id="t103" type="close" lemma="of" pos="P" morphofeat="IN">
     <span>
       <target id="w103"/>
     </span>
   </term>
<!--cake-->
   <term id="t104" type="open" lemma="cake" pos="N" morphofeat="NN">
     <span>
       <target id="w104"/>
     </span>
   </term>
```

Fig. 3. Lemmatisation and PoS tagging example for *Little Red Riding Hood*



```
    <!--HOOD-->
    <term id="t7" type="close" lemma="HOOD" pos="N" morphofeat="NN">
      <span>
        <target id="w7"/>
      </span>
<externalReferences><externalRef resource="wn_en30g.bin64" reference="09305358-n" conf
idence="0.497862"/><externalRef resource="wn_en30g.bin64" reference="03531546-n" confi
dence="0.318474"/><externalRef resource="wn_en30g.bin64" reference="10184081-n" confid
ence="0.0804221"/><externalRef resource="wn_en30g.bin64" reference="03531808-n" confid
ence="0.0676563"/><externalRef resource="wn_en30g.bin64" reference="08225334-n" confid
ence="0.0355862"/></externalReferences></term>
    <!--]-->
```

Fig. 4. WSD analysis for the lemma *hood*

4.1 Word clouds: working with global comprehension and vocabulary

Our first experiment focuses on the pre-reading stage, in which we aim at enhancing text comprehension by going through the plot of the story or the characters. More precisely, we have created word clouds which visually highlight the main ideas of the texts. Word clouds are visual representations of the words in a text and are typically used to depict the keywords. Commonly, most frequent words are displayed in bigger sizes, and thus give a straightforward insight on the topic of the text.

In fact, frequency of words shows a lot of potential in order to sketch the information in text, as high-frequency items cover a large proportion of words in text.

Hence, they have been worthy of attention by both language teachers and learners [23]. Furthermore, word clouds are a widely considered teaching resource [6, 12, 21, 29].

Taking all that into account, word clouds are convenient for a first introduction of texts as they offer a global view of the plot and can be a good tool to deal with the most relevant or specific vocabulary from a visual and playful approach. Additionally, word clouds can also be useful to compare two different texts, two authors on the same topic, namely.

In what regards the word cloud generation system, we have developed a prototype of a word cloud generator by combining the packages `Matplotlib` and `Numpy` in the Python programming language. All words from the original text are shown in the word cloud, except digits, punctuation marks and the so-called *stop words* (prepositions, determiners, conjunctions, etc.) that do not convey relevant information on the topic, in order to focus on meaningful words. We generate word clouds with the shape of an input image relevant to the tale we want to deal with so as to make the word clouds more appealing to the language learner.

In Figure 5 we present the final word cloud we have obtained from the *Little Red Riding Hood* tale. As can be seen, the word cloud acquires the shape of Little Red Riding Hood and the most frequent words are *grandmother, little, red-cap* and *good*. Dealing with those words is useful to start working with the vocabulary and the concepts the readers will find in the text.

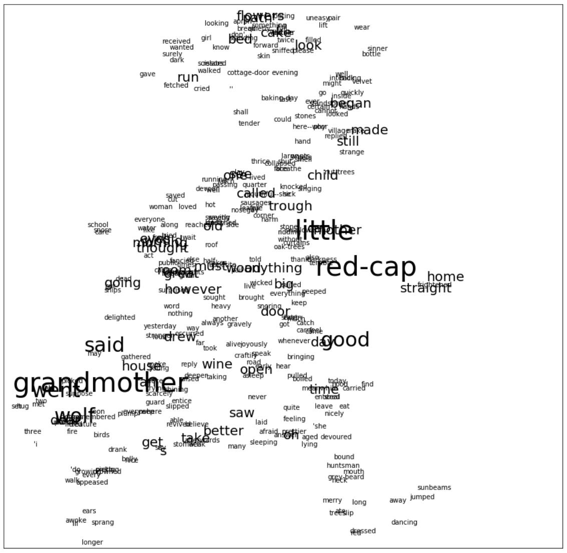


Fig. 5. Word cloud for *Little Red Riding Hood*

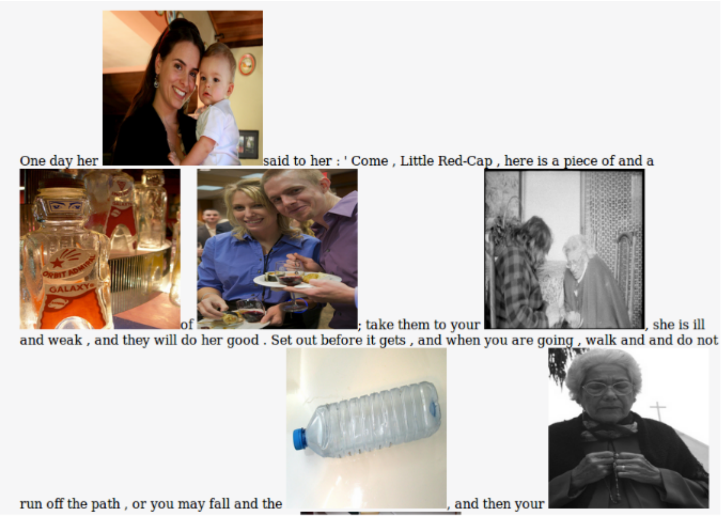
4.2 Pictotales: working on vocabulary with images

Pictotales are the tales in which nouns have been replaced with images from ImageNet. This is a single-language textual transformation approach in which some lemmas have been replaced for pictures.

In order to create the *Pictotales*, lemmas of some of the concrete nouns have been automatically looked up in ImageNet and an image corresponding to the lemma synset has been randomly chosen. In order to combine text and images, the narratives have been converted to HyperText Mark-up Language (HTML), which allows displaying text and images in web browsers and other visual interfaces.

As can be seen in Figure, 6, we have created a *Pictotale* from the English version of the *Little Red Riding Hood* tale. As one can see, concepts such as *mother*, *grandmother*, and *bottle* have been replaced with some relevant images.

This kind of exercises can be useful to help to learn vocabulary, using the images as contextualized flashcards or helping to evoke the target words in a first reading of the tales. Furthermore, *pictotales* can also be employed for performing vocabulary revision and memory exercises, naming the elements that appear in the images, all within the context of the tale.



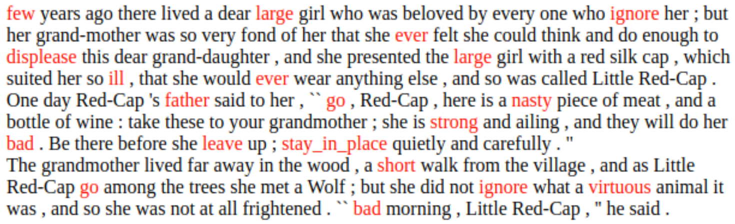
One day her said to her : ' Come , Little Red-Cap , here is a piece of and a of ; take them to your , she is ill and weak , and they will do her good . Set out before it gets , and when you are going , walk and and do not run off the path , or you may fall and the , and then your

Fig. 6. *Little Red Riding Hood* with images

4.3 Story revolution: working on vocabulary with related words

In the Story Revolution experiment, we have aimed at vocabulary learning through meaning-based substitutions. That is to say, we have substituted nouns and adjectives in texts with their antonyms, hyponyms or hypernyms.

Since we work with lemmatised texts, we have used the WordNet information for those. As WordNet can be understood as a net of semantically-based relations in which concepts are arranged from the most generic to the most specific and opposition relations are also included, obtaining the antonyms, hyponyms or hypernyms of the selected lemmas in text has been a straightforward process. Figure 7 displays a piece of *Little Red Riding Hood* where some words have been replaced with their antonyms (in red) e.g. *few, large, ignore....*



few years ago there lived a dear large girl who was beloved by every one who ignore her ; but her grand-mother was so very fond of her that she ever felt she could think and do enough to displease this dear grand-daughter , and she presented the large girl with a red silk cap , which suited her so ill , that she would ever wear anything else , and so was called Little Red-Cap .
One day Red-Cap 's father said to her , `` go , Red-Cap , here is a nasty piece of meat , and a bottle of wine : take these to your grandmother ; she is strong and ailing , and they will do her bad . Be there before she leave up ; stay_in_place quietly and carefully . "
The grandmother lived far away in the wood , a short walk from the village , and as Little Red-Cap go among the trees she met a Wolf ; but she did not ignore what a virtuous animal it was , and so she was not at all frightened . `` bad morning , Little Red-Cap , " he said .

Fig. 7. Passage of the English Version of *Little Red Riding Hood* with Antonyms

Replacing words with their hypernyms may help on text understanding for second language learners. In fact, removing the most obscure terms and offering more generic alternatives may lead, at least, to getting the global sense of the texts. We present a sentence of Little Red Riding Hood's tale in (1). As it can be seen, some of the words have been highlighted in different colours. In example (2), we present a sentence formed by the hypernyms of those highlighted words. In this second sentence more generic vocabulary is used and the sentence could presumably be better understood by non-proficient speakers of English.

(1) When the wolf had appeased his appetite.

(2) When the canine had calmed his craving.

On the contrary, when substituting words with their hyponyms, we help enlarging the available vocabulary and learning more specific words. In example (3) we present the outcome of replacing the highlighted words in (1) with their hyponyms.

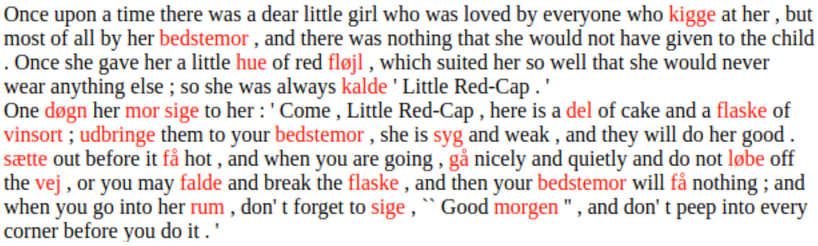
(3) When the coyote bear had appeased his stomach.

Word substitution is a common technique in lexical substitution and text simplification tasks, but generally word replacements are done with synonyms [10, 27, 15]. Nonetheless, hypernym and hyponym substitution can also be suitable for the task, since replacing all the words referring to the same concept for a single term makes us convey with one of the easy reading principles: "use the same term consistently for a specific thought or object" [25].

4.4 Uncovering words: discovering unknown words

In this experiment we wanted to automatically create one of the typical vocabulary learning exercises. Given a text in the target language, we have substituted some words for their translations in the learner's tongue. However, we have not opted for the individual translations of the words, but we have used WordNet to assign the term, in order to guarantee the terms share the same meaning.

For example, in Figure 8 we have taken as a basis the English version of the *Little Red Riding Hood* tale and we have translated some words into Danish (in red, *bedstemor, rum...*). Consequently, we have obtained a traditional "fill in the gaps" exercise with mother tongue clues.



Once upon a time there was a dear little girl who was loved by everyone who kigge at her , but most of all by her bedstemor , and there was nothing that she would not have given to the child . Once she gave her a little hue of red fløjl , which suited her so well that she would never wear anything else ; so she was always kalde ' Little Red-Cap . '
One døgn her mor sige to her : ' Come , Little Red-Cap , here is a del of cake and a flaske of vinsort ; udbringe them to your bedstemor , she is syg and weak , and they will do her good . sætte out before it få hot , and when you are going , gå nicely and quietly and do not løbe off the vej , or you may falde and break the flaske , and then your bedstemor will få nothing ; and when you go into her rum , don' t forget to sige , `` Good morgen " , and don' t peep into every corner before you do it . '

Fig. 8. Passage of *Little Red Riding Hood* in English with some words in Danish

Conversely, this technique can be applied in order to present the text in a language that the learner speaks and translate several words into the target language. For example, the text in Figure 8 could be used by English speakers learning Danish. In this way, learners find a comfortable context in which they can focus on the vocabulary. One possible exercise that can be done is encouraging the learners to make hypothesis about what a word means given its surrounding context in order to uncover the story in text.

5 Discussion

In this paper we have presented four text transformations that not only can help teachers to create exercises but also writers or editors to create new texts. In fact, our approach may shorten the exercise generation time and can also help the creativity of the professionals, namely when looking for a convenient translation. Nevertheless, we have to underline that these texts need a revision before they are used. In particular, we detail next three main types of shortcomings.

1. Sometimes the absence of a linguistic item prevents from addressing the transformation query to generate a variant in the text. An illustration of this can be seen in examples (1) and (3) in Section 4.3 where the same form (*had appeased*) is offered for both the target term and its hyponym (autohyponymy). According to the resources used, there is not a more specific way of referring to the event, and probably it will also be difficult for an expert to come up with one.
2. Additionally, as pictures from ImageNet are chosen randomly among all the images linked to a certain synset, the picture selected might not be the most suitable according to the context. For example, in Figure 6 the word *bottle* is represented by an empty plastic bottle, but could be better represented by a full

glass bottle of wine. We reckon an optimal approach to *pictotale* generation would be a system that takes into account the whole narrative context for image selection.

3. Some other issues, instead, arise from processing errors. In some cases, the incorrect Part-of-Speech tag assignation may lead to the substitution of a wrong word. In other cases, incorrect word sense disambiguation may be a source of errors. For example, in (4) we present a passage of the *Little Red Hood* in Spanish where some words have been translated into Danish (as when uncovering words, Section 4.4). As depicted in the example, we have found that the Spanish word *chica* (girl) has been substituted with the Danish word *dreng* (boy). This error seems due to the fact that the Spanish lemmatiser gives the masculine form *chico* (boy) as lemma. UKB disambiguates the masculine lemma as *boy* and our script relies on UKB's disambiguation to look up the Danish word.

(4)	(...)	nada	que	no	le	hubiera	give	a	la	dreng	(...)
	(...)	nothing	that	not	her	had.SUBJ	give	to	the.FEM	boy	(...)

'(...) nothing that hadn't been *given* to the(fem) *boy*.'

Moreover, in case of the verbs, as our tools substitutes lemmas, it is necessary to fix the conjugation. An example of this is the Danish verb *give* (give) in (4), which should be corrected to the participle *givet* (given). In order to overcome this problem, natural language generation techniques that take the syntax of the target language into account should be used. Nonetheless, offering only the lemma does not invalidate the proposed exercise as guessing or generating the right verbal tense from the context is also a possible exercise.

Despite the shortcomings, this method offers great capacity for text adaptations. In this work we have applied our modifications to all the words, but that can be easily customised. Possible customisations are substituting less frequent words, longer words, complex words and keywords/keyphrases among others.

For example, regarding frequent and infrequent words, word frequency lists such as Ogden's Basic English word list or the ones that take word distributions into account [7] can be employed in order to set the threshold of interest (very frequent, frequent, normal, infrequent, not frequent).

In what concerns word length, the Plain Language guidelines the use short words is recommended (a summary of guidelines can be found in [22]). Shorter words can be easier to learn; however, the influence of word length does not seem to be the only factor for our memory [17].

These text adaptation processes may also be enhanced with other current NLP techniques, such as complex word identification [24]. In this preprocessing step for lexical simplification, complex words and expressions are identified in order to replace them later with simpler equivalent alternatives [26]. Our approach may be integrated in this step so as to replace complex words and expressions with simpler equivalent alternatives.

Besides, keyphrase detection, which deals with finding most important words in texts, is another NLP task that has been approached in the context of information extraction [20, 28] but that can easily help the creation of text adaptations for students with special needs, by helping the readers to get the main ideas of the text.

Taking all that into account, we reckon that, although we have centred our experimentation in vocabulary learning exercise generation and that our system admits some improvements, the tools and resources proposed in our experimentation can be of much help in other DH research trends in which text transformations play a significant role. Additionally, even if we have presented our transformation proposals as isolated experiments, they can all be combined to address specific needs.

6 Conclusion and Future Works

We have conducted a series of experiments in which we have altered texts through NLP methods in order to create resources for vocabulary learning. Exactly, text adaptations have been i) creating word clouds, ii) texts with images, iii) texts with different but semantically related works and iv) texts with translations. Our main objective was to show a possible support to teachers/educators when creating vocabulary learning or reading comprehension exercises by means of NLP applications. This approach can also be seen as a computational creativity exercise as transformations and suggestions have been automatically generated.

Our experiments offer just the first insights on what NLP can offer for textual modification. Automatic methods are far from being perfect and still need human supervision, but it is undeniable the fact that they ease the burden of coming up with suitable ideas in certain contexts. Further, taking into account that our approach is a NLP-based one, we consider that in the next steps conducting both qualitative and quantitative evaluations in real scenarios is fundamental so that to measure the actual performance of our implementations and to integrate our preliminary proposal within the scientific framework.

In the case of educational purposes, a first step on improving our work should be evaluating the text adaptations with target audiences in order to better understand their needs; e.g. which words should be adapted or what new approaches they may require. That is why we encourage the collaboration with other experts in the area of humanities and, specially, within the domain of education.

Moreover, we think our proposal can be adapted to address more education needs. We foresee the following applications: i) adapting the texts with pictograms, ii) going further than words and substituting phrases iii) creating games removing words from texts or giving their definitions in order to guess them, or iv) giving two words and guessing their relation. Creating an online text adaptation application, where each one can customise its text, is also one of our future goals.

Out of the education domain, we believe our approach may also have different uses. For example, we think it might be useful for museums or other cultural institutions for the creation of adapted or multimedia and interactive material. In the case of the artistic creation of word clouds automatically extracting the most relevant words can also be of great help. Finally, we want to reinforce the idea that using NLP methods might impulse computational creativity for new kinds artistic expressions.

7 Acknowledgements

We thank Codefest summer school for providing the infrastructure to begin this work. We also thank Larraitz Uria for her contribution to the early steps of this work.

References

1. Agerri, R., Bermúdez, J., Rigau, G.: IXA pipeline: Efficient and Ready to Use Multilingual NLP Tools. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 3823–3828. European Language Resources Association (ELRA), Reykjavik, Iceland (2014)
2. Agirre, E., Edmonds, P.: Word sense disambiguation: Algorithms and applications, vol. 33. Springer Science & Business Media (2007)
3. Agirre, E., López de Lacalle, O., Soroa, A.: The risk of sub-optimal use of Open Source NLP Software: UKB is inadvertently state-of-the-art in knowledge-based WSD. In: Proceedings of Workshop for NLP Open Source Software (NLP-OSS). pp. 29–33. Association for Computational Linguistics (2018), http://aclweb.org/anthology/W18-2505
4. Agirrezabal, M., Arrieta, B., Astigarraga, A., Hulden, M.: POS-Tag Based Poetry Generation with WordNet. In: 14th European Workshop on Natural Language Generation (ACL 2013). pp. 162–166 (2013)
5. Alhawiti, K.M.: Natural Language Processing and its Use in Education. International Journal of Advanced Computer Science and Applications **5**(12), 72–76 (2014), http://thesai.org/Downloads/Volume5No12/Paper_10-Natural_Language_Processing.pdf
6. Baralt, M., Pennestri, S., Selvandin, M.: Using wordles to teach foreign language writing. Language Learning & Technology **15**(2), 12–22 (2011)
7. Bentz, C., Alikaniotis, D., Samardžić, T., Buttery, P.: Variation in word frequency distributions: Definitions, measures and implications for a corpus-based language typology. Journal of Quantitative Linguistics **24**(2-3), 128–162 (2017)
8. Bird, S.: NLTK: the Natural Language Toolkit. In: Proceedings of the COLING/ACL on Interactive presentation sessions. pp. 69–72. Association for Computational Linguistics (2006)
9. Bond, F., Foster, R.: Linking and Extending an Open Multilingual Wordnet. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1352–1362. Association for Computational Linguistics (2013), http://aclweb.org/anthology/P13-1133
10. Bott, S., Rello, L., Drndarevic, B., Saggion, H.: Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In: Proceedings of COLING 2012: Technical Papers. pp. 357–374 (2012)
11. Carreras, X., Chao, I., Padró, L., Padró, M.: FreeLing: An Open-Source Suite of Language Analyzers. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). pp. 239–242 (2004)
12. Dalton, B., Grisham, D.L.: eVoc Strategies: 10 Ways to Use Technology to Build Vocabulary. The reading teacher **64**(5), 306–317 (2011)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A Large-scale Hierarchical Image Database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009). https://doi.org/10.1109/CVPR.2009.5206848
14. Fellbaum, C.: Wordnet: An Electronic Lexical Database. MIT Press Cambridge (1998)

15. Ferrés, D., Saggion, H., Guinovart, X.G.: An Adaptable Lexical Simplification Architecture for Major Ibero-Romance Languages. In: Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems. pp. 40–47 (2017)
16. Gonçalo Oliveira, H.: A Survey on Intelligent Poetry Generation: Languages, Features, Techniques, Reutilisation and Evaluation. In: Proceedings of the 10th International Conference on Natural Language Generation. pp. 11–20. Association for Computational Linguistics (2017), http://aclweb.org/anthology/W17-3502
17. Guitard, D., Gabel, A.J., Saint-Aubin, J., Surprenant, A.M., Neath, I.: Word Length, Set Size, and Lexical Factors: Re-examining what Causes the Word Length Effect. Journal of Experimental Psychology: Learning, Memory, and Cognition **44**(11), 1824 (2018)
18. Litman, D.: Natural Language Processing for Enhancing Teaching and Learning. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16). pp. 4170–4176. Association for the Advancement of Artificial Intelligence (2016)
19. Lwin, S.: Using Folktales for Language Teaching. The English Teacher **44**(2), 74–83 (2015)
20. Mahata, D., Kuriakose, J., Shah, R.R., Zimmermann, R.: Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 634–639. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/N18-2100, http://aclweb.org/anthology/N18-2100
21. McNaught, C., Lam, P.: Using Wordle as a supplementary research tool. The qualitative report **15**(3), 630–643 (2010)
22. Mitkov, R., Štajner, S.: The fewer, the better? A Contrastive Study about Ways to Simplify. In: Proceedings of the Workshop on Automatic Text Simplification-Methods and Applications in the Multilingual Society (ATS-MA 2014). pp. 30–40 (2014)
23. Nation, I.S.: Learning vocabulary in another language. Ernst Klett Sprachen (2001)
24. Paetzold, G., Specia, L.: Semeval 2016 Task 11: Complex Word Identification. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 560–569 (2016)
25. PlainLanguage.gov: Federal Plain Language Guidelines (2011), https://plainlanguage.gov/media/FederalPLGuidelines.pdf
26. Shardlow, M.: Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014). pp. 1583–1590 (2014)
27. Specia, L., Jauhar, S.K., Mihalcea, R.: Semeval-2012 task 1: English Lexical Simplification. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. pp. 347–355. Association for Computational Linguistics (2012)
28. Wang, R., Liu, W., McDonald, C.: Using Word Embeddings to Enhance Keyword Identification for Scientific Publications. In: Australasian Database Conference. pp. 257–268. Springer (2015)
29. Williams, W., Parkes, E.L., Davies, P.: Wordle: A method for analysing MBA student induction experience. The International Journal of Management Education **11**(1), 44–53 (2013)