Corpus-based online word formation exercises for advanced learners of English – challenges and solutions

Grzegorz Krynicki

Faculty of English, Adam Mickiewicz University, Poznań, Poland

Abstract. The paper presents the design and operation of an online platform for word formation practice. The system is based on a pre-defined list of pairs of base and derived forms and usage examples drawn automatically from the British National Corpus. A procedure for the extraction of example sentences is outlined. Results of 372 users' interacting with the system for over 4.5 months are reviewed. The question about what factors influence users' evaluation of specific exercises as more difficult is addressed. The results may be relevant in the area of language testing, preparation of examination materials, student-teacher online interaction and teaching English word formation.

1 Word formation in learning English as a foreign language

Advanced learners of English willing to expand their vocabulary appreciate the study and practice of word formation. Knowing how to combine a small set of particles (prefixes like *non-*, *im-*, *de-* or suffixes like *-able*, *-ish*, *-ly*) with a few base words may increase learner's vocabulary significantly and with minimum effort. For example, knowing what these particles mean and the meaning of the simple root morpheme *port* ('to send' or 'carry') most advanced learners would also probably guess the meanings of *export*, *import*, *deport*, *portable* and *transport*. To look at it from another perspective – in the British National Corpus (BNC) the prefix over- begins 2013 different word types (Joandi 2012: 13). Knowing how over- influences the meaning of words it is attached to allows to know close to half of what each of these 2013 words means.

The set of corpus-based gap fill exercises in word formation described in this paper is based two major sources, a word formation list of 1929 tokens (Szczegóła 2012) and the British National Corpus of 100 million tokens. The word formation list was prepared for 1-3BA and 1MA students of the Faculty of English, Adam Mickiewicz University in Poznań, Poland by teachers of practical English. However, they can be useful to all advanced learners of English (from B2 to C2 in CEFR). The list of word forms was compiled based on articles and course-books used by students of English philology. They illustrate most word-formation mechanisms (prefixation, suffixation, compounding, clipping etc.) and cover a wide range of general topics. All example sentences were drawn from the BNC (BNC 2001) to ensure that the language used in these activities is authentic and varied.

These exercises are aimed at advanced students of English who want to:

- improve their receptive and productive command of English vocabulary,
- inductively learn English word formation rules,
- overcome the interference from their native tongue morphology,
- master vocabulary in authentic sentence context,
- learn Polish equivalents of English complex words (although the knowledge of Polish is not necessary to benefit from all other aspects of the exercises),
- practice for advanced English grammar tests and examinations,
- improve their skills in dictionary word lookup – dictionaries often provide definitions for simpler words and leave the creation of complex words with relatively intuitive meanings to the user.

The system may also be used by EFL teachers and test designers. Although new corpora and other lists of base form – derived form pairs can easily be added by the administrator to create new exercises, the online interface does not allow manipulating these resources. Free unrestricted online access to the exercises is possible at the following address: http://wa.amu.edu.pl/~krynicki/wf

2 How to use the system

When the user logs into the system, he will see a table of 6 columns (Fig. 10, last page of this paper). In the 2nd column of the table, 10 example sentences are listed, each with a gap that needs to be filled with a word form derived from the base word given in the 3rd column. If the example sentence is too ambiguous, the user may click "More" to see additional example sentences. If are ready to see the answer, click "Answer" in the 4th column. The user compares his answer with the answer that appears in the 5th column and mark check-box in the 6th column if the user's answer differed in any way from the answer provided by the system. Once the user has done all the 1929 exercises, he will have the possibility to export the difficult items to a tab-separated text file so that he can drill them in spaced memory software, e.g. Anki (2013) or SuperMemo (2013).

All examples were drawn from the corpus automatically so it may happen that even top students will have problems guessing the missing word form on the basis of a single ambiguous example sentence. For this reason, the option of viewing two additional example sentences has been provided. If the user clicks "More" – a new sentence will drop down below the already visible example. If the exercises are used to practice for a written examination, it should be kept in mind that in most exams where word form gap fill exercises appear the user will not have the possibility to see more than 1 example sentence. Moreover, the user will have to write his answers not just think about them as is the case with this system. For these reasons, before providing the answer, the user should try to mentally spell the word and mark it as difficult if he makes the slightest mistake.

If the user does not know what the English word form means in Polish, a list of equivalents will appear in a balloon tip when the user hovers his mouse pointer over most word forms. If the word form is clicked, the user will be redirected to a form where he can edit the Polish equivalents of the word form and English example sentences. The editions will be visible to others after they have been accepted by the administrator. In the system, Polish equivalents were drawn automatically from various electronic English-Polish dictionaries without any regard to their part of speech (POS), order in which they originally appeared or phrases they may be used in.

Each student had a different order of sentences submitted to him. The order was generated in a pseudo-random fashion during his first visit. Randomization was adopted as a precaution against students who would like to solve the exercises simultaneously on different computers and help each other. Every time the student logs into the system he can continue his work without having to repeat the exercises he has already done.

At the bottom of the screen the user sees the progress bar so that he can monitor how many exercises out of 1929 he has done.

3 Selection of example sentences

Automatic selection of example sentences was conducted taking into consideration the length of the candidate sentences and the number of proper names they contained.

Roughly, the more the example sentence approached the "ideal" length and the fewer proper names it had, the more chances it had of being selected.

1. The BNC corpus was split into approx. 6 million sentences.
2. Corpus entities were converted to Windows-1252 encoded text to make their tokenization and display easier, e.g. the entity *&bquo;* used in BNC to denote a double quotation mark is not a standard HTML entity and was converted to *"* .
3. Tokenization and down-casing, e.g. *She can't stand her mom's "complaints".* was converted to *she can not stand her mom 's " complaints " . .*
4. By the rule of the thumb
    - Sentences of 80 characters or fewer were excluded as they were considered to provide not enough context to guess the gapped word. Although excessively long sentences often contain material irrelevant for the guessing of the gapped word, the upper limit for the sentence length was not set. The ideal sentence length was set at 160 characters;
    - Sentences containing a capital letter anywhere else than at the beginning of the sentence were excluded in the first stage to minimise the number of proper names and abbreviations in the example sentences.
5. For each of the remaining sentences:
    - Base form of each word in the sentence was obtained by consulting lemmatized word frequency lists (Kilgarriff 1995);
    - If the base form was present in the WA list (among lower-case derived words), the sentence was considered a potential example of the usage of this base form;
    - Potential examples were ordered from the ones closest to the ideal length to the ones farthest from the ideal length. In this order example sentences were submitted to the student.
6. If 3 sentences meeting the above criteria were not found for a word form from the WA list, in the second stage, the missing sentences were filled in from those containing capital letters elsewhere than at the beginning of the sentence in the increasing order of the number of capital characters they contained.

As an effect of this procedure, in 92.3% of exercises the word form was illustrated by 3 example sentences, in 4.5% of exercises 2 example sentences were used and 1 sentence was used to illustrate the usage of the remaining 3.2% of word forms.

No of sentences	*Frequency*	*Relative frequency*
1	62	0.0321
2	86	0.0446
3	1781	0.9233

Table 1: Frequency of exercises with 1, 2 or 3 example sentences.

The author is aware of many imperfections the above algorithm has, especially in the view of solutions proposed by e.g. Kilgarriff 2008 or Didakowski at al. 2012. In future stages of the project, parameters that characterize the readability, complexity and stylistic properties of the examples will be considered.

4 Students' judgements about the difficulty of the exercises

Two freshman groups attending classes in FCE General English and CAE English Grammar were suggested to use the platform to prepare for their final practical English examination. A notice about the exercises was also published on a Moodle site devoted to practical English examination that all and only Faculty members had access to. The notice additionally informed that example sentences used in the exercises would not appear in the final exam. Students were also reminded that the word formation component of the exam will include only the words from the WA list.

The Faculty members included over 1500 BA and MA students from B2 to C2 CEFR levels. Over the period of 4 months and 20 days (Apr 9th – Aug 29th), 389 students logged into the system at least once. For 17 of them, there is no evidence of them doing any exercises as batches of more than 10 completed exercises were evaluated. The remaining 372 students did 417.3 out of of 1929 exercises on average (21.6%). 45 students completed all 1929 exercises. Each exercise was solved by at least one student. A unique exercise was solved by an average of 23.8 students.

In order to improve the interface and the content of the exercises as well as to aid the preparation of tasks for the final practical English examination, an analysis of the students' responses was conducted. Students' responses included information about which exercises they found difficult. The difficulty judgements were then related to properties of the prompt base word, expected word form and the example sentence from which it was extracted.

4.1 Statistics on students' judgements

The user of the system was encouraged to mark the exercise as difficult if he made any mistake in it. He was informed that once all the exercises have been completed, difficult items could be exported for drilling in spaced memory software. 254 out of 372 students (68.3%) marked at least one exercise as difficult. Among these students, the average number of items marked as difficult was 109.0 i.e. 16.2% (standard deviation of 231.2). This constituted 26.1% of all the exercises they tried to solve on average. The maximum percentage of exercises a single student marked as difficult 68.2% (225 of 330). 1919 exercises out of 1929 were marked as difficult at least once. Fig. 2 illustrates the distribution of exercises with different levels of difficulty according to students.

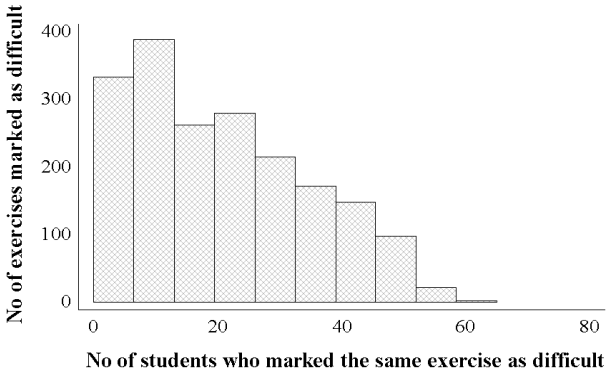

Fig. 1: Histogram of 1919 exercises marked difficult at least once by 254 students with the numbers of times one exercise was marked difficult and observations pooled into 10 classes.

Students found the exercises rather challenging. One exercise was marked difficult by 61 students. Half of the exercises were marked difficult 19 or more times. 36 exercises were marked difficult exactly once.

4.2 Factors potentially influencing students' judgements

In this study, the influence of the following factors on users' evaluation of exercise difficulty was considered: student's language competence (for a sample of students), properties of base form hint and the gapped word form, number of example sentences illustrating the use of the word form and their length as well as availability of Polish translation of the word form. The potentially significant factors that were not considered include student's language aptitude and the properties of the context in which word form appeared were not considered.

Properties of the base word and derived form included their frequencies in BNC frequency list (Kilgarriff 1995), similarity of POS codes listed for them in the BNC list as well as their graphemic similarity.

To obtain these properties for base and derived forms, the BNC list was preprocessed in the following way:

- Complex tags were “rounded” to general categories of nouns, adjectives, adverbs, verbs (e.g. NN0 common noun and NN1 singular common noun were pooled into NN category). All other POS were discarded;
- From portmanteau tags, used in CLAWS to indicate where the system was uncertain between two possible analyses, only the first one was chosen;
- Frequencies of identical words within the same rounded and simplified POS were added.

The resulting list contained 365040 nouns, 68460 verbs, 190086 adjectives and 9739 adverbs.

Graphemic similarity between base form and word form was another characteristic whose influence on exercise difficulty was considered. Graphemic similarity was expressed by two parameters. First, by the longest common prefix between the two forms. It was calculated as the maximum number of characters that the two words shared at their beginnings. Second, by the longest common subsequence ratio (LCSR) determined by dividing the length of their longest common subsequence by the length of the longer word (Melamed 1999). It was hypothesized that the greater the similarity between the two forms, the easier is will be to guess the derived from given the base form and the less frequently an exercise including them will be marked as difficult. Table 2 includes an extract from the list of 1929 word forms annotated for similarity and frequency information (for complete list refer to Krynicki 2013b).

4.3 Significance tests of factors potentially influencing students' judgements

Statistical significance tests were used to identify factors that had a significant influence over users' judgements about exercise difficulty. The dependent variable in all tests was the number of times a given exercise was judged difficult (ranging from 0 to 61). The independent variables of word frequency were grouped under 4 or 5 variables: 0 if 0 frequency was observed, 1-3 for data points up to 25th, 50th and 75th percentile respectively and 4 otherwise (Table 3). Longest common subsequence ratio, longest common prefix lengths (“prefix” in a pattern-matching rather than linguistic sense) and length of the first sentence were grouped as presented in Table 4.

Base	*POS*	*Base freq*	*Word form*	*WF POS*	*WF freq*	*Intersection*	*Long. Prefix*	*LCSR*	*Pol. equ.*	*No sent.*
abandon	nv	1316	abandonment	n	496	n	7	0.64	1	3
able	j	30410	ability	n	9135		2	0.43	1	3
normal	jd	12452	abnormal	j	810	j	0	0.75	1	3
normal	jd	12452	abnormality	n	287		0	0.55	1	3
normal	jd	12452	abnormally	d	151	d	0	0.6	1	3
cite	v	282	above-cited		151		0	0.6	0	3
abstain	jnv	129	abstainer	n	3	n	7	0.78	1	3
abstain	jnv	129	abstention	n	99	n	4	0.6	1	3
abstain	jnv	129	abstinence	n	150	n	4	0.6	1	3
abstain	jnv	129	abstinent	j	10	j	4	0.67	1	2

Table 2: First 10 word forms annotated for base form, parts of speech, frequencies in BNC, intersection of the sets of POS tags for each form, length of the longest common prefix, longest common subsequence ratio (LCSR), information about whether the option of displaying Polish translation of the word form was available, number of example sentences that illustrated the use of the word form. POS abbreviations: j – adjective, d – adverb, v – verb, n – noun. The whole list is available at http://wa.amu.edu.pl/~krynicki/wf/table2.csv

Grouping variable	*Range*	*Freq of base form (x)*	*Frequency of derived form (x)*
1	$0 < x \leq 25\%$	5-1658	1-68
2	$25\% < x \leq 50\%$	1659-4372	69-273
3	$50\% < x \leq 75\%$	4373-11650	274-1062
4	$75\% < x \leq$ max	11651-129547	1063-48374

Table 3: Transformation of Frequency of base form *in BNC and* Frequency *of derived form into 4 grouping variables*.

Grouping variable	*LCSR*	*Longest common prefix*	*Length of the first sentence*
0	0-0.15	0-2	52-110
1	0.16-0.50	3-4	111-159
2	0.51-0.60	5	160
3	0.61-0.67	6	161-200
4	0.68-1	7-11	201-391

Table 4: *Transformation of* Longest common subsequence ratio, Longest common prefix *and* Length of the first sentence *into 5 grouping variables.*

In all tests at least one of ANOVA assumptions was violated – the standardized skewness and/or kurtosis was outside the range of -2 to +2 for at least one of the factors and/or the difference between the smallest standard deviation and the largest

was greater than 3 to 1. Therefore, Kruskal-Wallis Test (KWT) was used to test significance of most factors.

Language competence

General English written test results were known for 21 of 254 students who marked at least one exercise as difficult. A relatively weak positive correlation was found between student's results and the number of exercises he marked difficult (Spearman rank correlation coefficient.= 0.22, p = 0.023). This indicates that marking exercises was not directly related to language competence but it may have rather reflected student's willingness to review items in the future to remember them better in the practical English exam, student's diligence in general or student's preference for reviewing items in spaced memory software rather than using web interface.

Base form frequency

KWT was used to test the null hypothesis that the medians of grouping variable of *Times judged difficult* (i.e. how many times an exercise was marked difficult by all students) within each of the 4 levels of the grouping variable of *Frequency of base word* are the same. The test statistic K = 10.15 and p = 0.0173, which is a significant result at the 0.05 level. Fig. 2 presents a Box-and-Whisker plot of the dependent variables against the factor. Boxes extend to 1st and 3rd quartile, whiskers extend to the maximum observations. Notches that do not overlap indicate medians that are significantly different. Therefore, contrasts between the levels 1:2, 1:3, 2:4 and 3:4 are statistically significant. In other words, exercises using most and least frequent base forms as hints are judged significantly more difficult than those using hints of frequency between 25th and 75th percentile.

Possible reasons may be related to higher derivational productivity of most frequent base forms and their higher ambiguity. Low-frequency base forms may be of less help as a hint because of their lower familiarity to students.

Word form frequency

The aim of the second test was to test the null hypothesis that the medians of Times judged difficult within each of the 4 levels of Frequency of derived form are the same. The test statistic K = 15.1029, p = 0.0017, which is significant at 0.05 level.

Derived forms of low frequency were difficult to guess if gapped from an exercise probably because of their low familiarity to students.

Scalar of intersection of POS tag sets for base and derived forms

Consider two sets, one containing POS tags listed in BNC for base word used as a hint in our word formation exercise and the other containing POS tags for the form derived from the hint but gapped in the example sentences. Intersection of these two sets is the set of POS tags base and derived forms have in common. The scalar (or cardinal) of the intersection is the number of POS tags shared by both forms. Fig. 4 illustrates the result of KWT of *Times judged difficult* against the *POS intersection scalar.*

This effect to some extent may be explained by the fact that students assume derivation usually changes morphosyntactic category of the base form. Therefore, the greater the overlap between POS tags of the two forms, the more problematic such

derivation may appear. This effect is also reinforced by the fact that the greater the intersection, the greater the POS set of each form and the greater their ambiguity.

Longest common prefix and LCSR

The first level of *Longest common prefix* (0 indicating prefixes of 0-2 characters) differs significantly from all the other levels (Fig. 5) with respect to the difficulty of exercises containing forms that begin with this prefix.

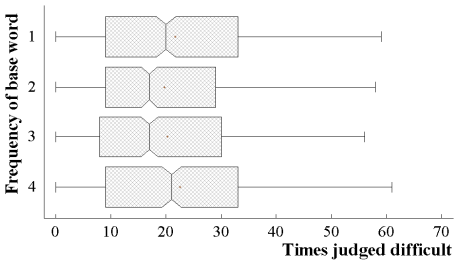

Fig. 2: Times judged difficult vs. Frequency of base word. K = 10.15, p = 0.0173

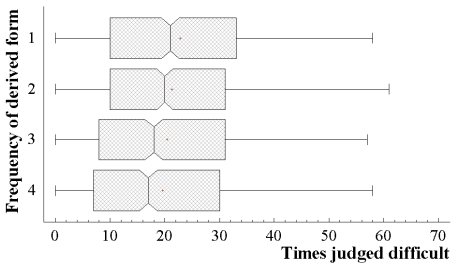

Fig. 3: Times judged difficult vs. Frequency of derived form. K = 15.1029, p = 0.0017

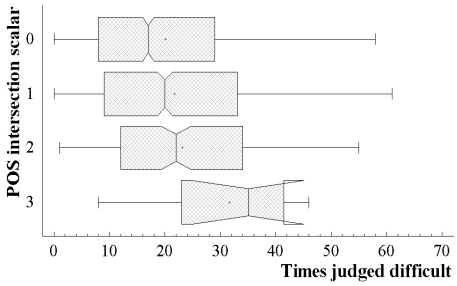

Fig. 4: Times judged difficult vs. POS intersection scalar. K = 15.3427, p = 0.0015

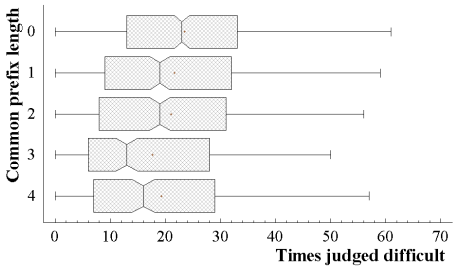

Fig. 5: Times judged difficult vs. Common prefix length. K = 49.625, p = 0.0000

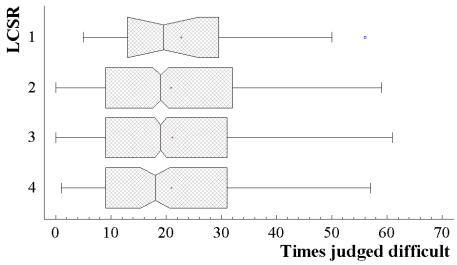

Fig. 6: Times judged difficult vs. LCSR. K =0.3798, p = 0.9443

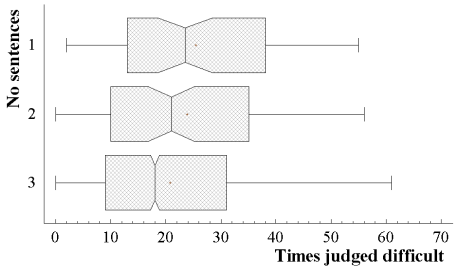

Fig. 7: Times judged difficult vs. Number of example sentences. K = 9.20819, p = 0.0100

Short common prefix or lack of it may make exercise more difficult. Similarity anywhere within the words as expressed by LCSR (Fig. 6) does not significantly affect how exercises containing them are evaluated ($p = 0.9443$).

Number of example sentences and sentence length

Students' judgements indicate that having 3 example sentences made their task significantly easier than when they have just 1 example (Fig. 7). It is also possible however that the greater difficulty of exercises with 1 example sentence may follow from the fact that if only 1 usage example meeting criteria described in 3 was found in BNC for the given word form it must be rare and therefore difficult no matter how many examples it would be illustrated with.

Sentences shorter than 111 characters as well as those longer than 200 increase the chances that the student will find the exercise difficult (Fig. 8). This last result may have been reinforced by time pressure before the exams – reading lengthy sentences may have been considered by students a waste of time.

Polish equivalents

After trying to guess the English derived form, the user could look up the correct answer in English and make sure he knew its Polish equivalents. Learning new Polish meanings could influence his decision about whether to mark the exercise as difficult. KWT revealed a significant relationship between the presence of Polish equivalent and the difficulty of the exercise (Fig. 9).

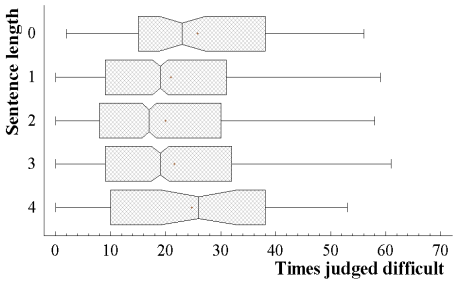

Fig. 8: Times judged difficult vs. Sentence length. K = 15.1209, p = 0.0045

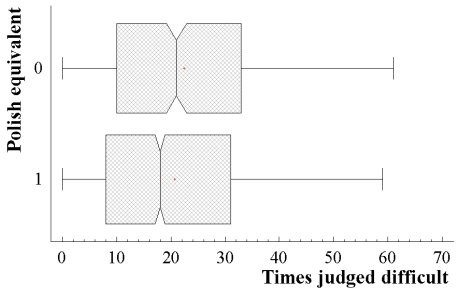

Fig. 9: Times judged difficult vs. Polish equivalent. K = 3.8961, p = 0.0484

5 Conclusions

The experiment described in this paper indicates that even a simple method of automatic selection of example sentences from a sufficiently large corpus may result in useful and engaging word formation exercises. The benefit of exercises of this form is not limited to acquiring correct English word-formation rules. Due to authentic sentence context, they develop the learner's grammar skills, teach meaning and meaning relationships and collocations. Moreover, a well designed word formation exercise with appropriate context is not only more effective but also more interesting than isolated word lists (c.f. Balteiro 2011: 28).

Practical conclusions that follow from the above study may include:

- Word forms derived from base forms by other processes than prefixation are considered more difficult and should probably be paid greater attention to by learners and teachers;
- Word formation exercises using most and least frequent base forms as hints are more challenging than hints of average frequency;
- Students should be aware that derivation does not always change morphosyntactic category of the base form;
- With automatically extracted examples, it is important that they have alternatives;
- The absence of L1 equivalents of gapped word forms increases the perception of the exercise as a difficult.

Future version of the system will incorporate methods of example sentence extraction so that the context of the gapped word form is balanced for frequency and so that important collocations of the word form are represented. Other forms of word formation exercises will also be introduced.

Bibliography

[1] Anki. 2013. Spaced repetition software. http://ankisrs.net

[2] Balteiro, Isabel. 2011. Awareness of L1 and L2 Word-formation Mechanisms for the Development of a More Autonomous L2 Learner. In: Porta Linguarum (15), pp.25-34.

[3] Didakowski, Jörg, Alexander Geyken and Lothar Lemnitzer. 2012. Automatic example sentence extraction for a contemporary German dictionary. In: Proceedings EURALEX 2012, Oslo, pp. 343-349.

[4] Joandi, Linnéa. 2012. Productivity Measurements Applied to Ten English Prefixes. http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-81966

[5] Kilgarriff, Adam. 1995. BNC database and word frequency lists. http://www.kilgarriff.co.uk/bnc-readme.html

[6] Kilgarriff, Adam, Miloš Husák, Katy McAdam, Michael Rundell and Pavel Rychlý. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In: Proceedings of the XIII EURALEX International Congress, pp. 425-432.

[7] Krynicki, Grzegorz. 2013a. Corpus-based word form gap fill exercises for advanced learners of English. http://wa.amu.edu.pl/~krynicki/wf/

[8] Krynicki, Grzegorz. 2013b. WA list of ~1929 word forms annotated for frequency and similarity information. http://wa.amu.edu.pl/~krynicki/wf/table2.csv

[9] Melamed, I. Dan. 1999. Bitext maps and alignment via pattern recognition. Computational Linguistics, 25(1): pp. 107–130.

[10] Supermemo. 2013. Spaced repetition software. http://www.supermemo.pl/

[11] Szczegóła, Tomasz. 2012. WA list of ~1929 word forms was compiled by and is available on WA UAM PNJA Moodle site: http://wa.amu.edu.pl/moodle/mod/resource/view.php?id=32196

[12] The British National Corpus, version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

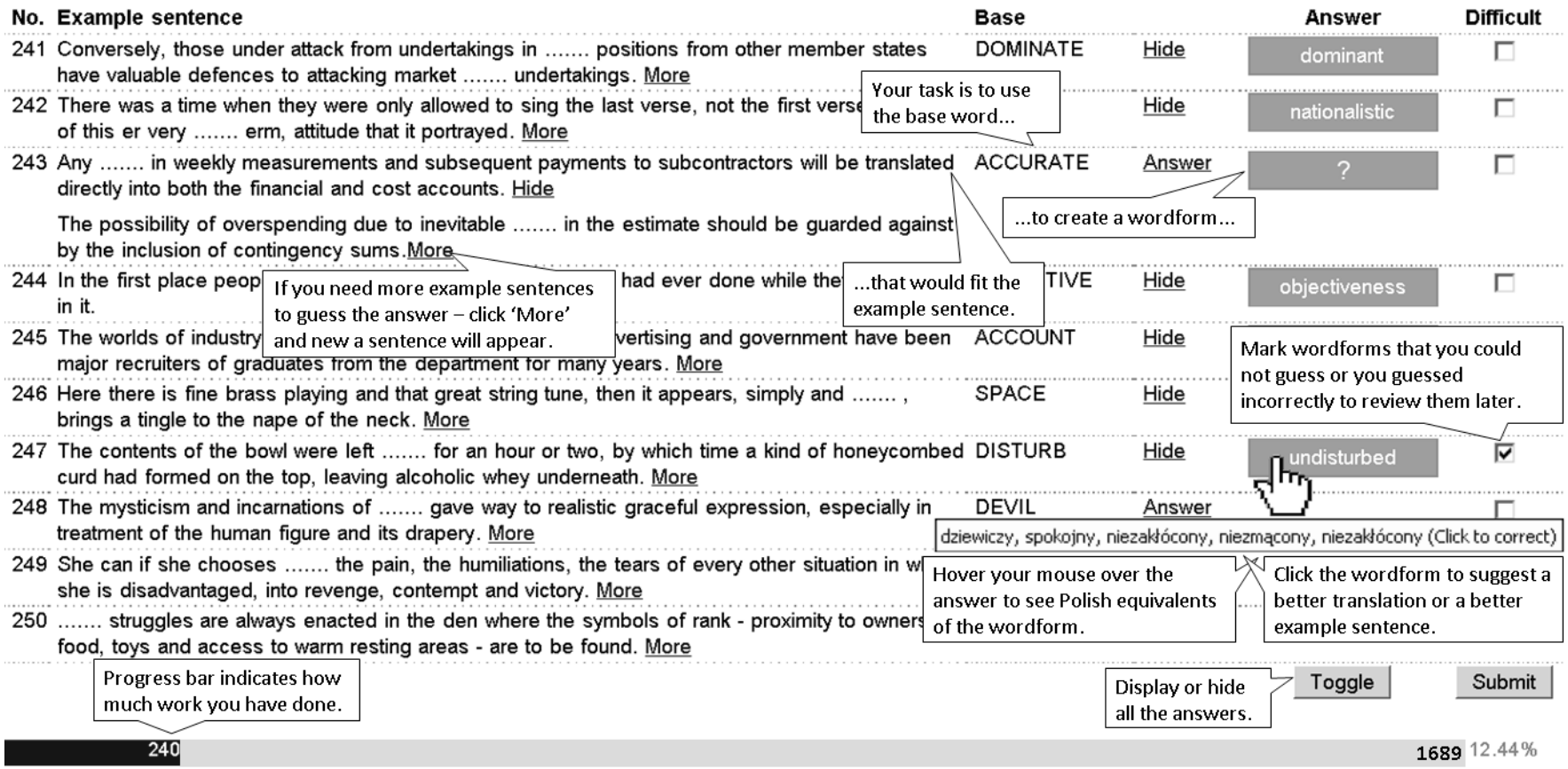

Fig. 10. A screenshot of the exercise panel with callouts explaining all functionalities.