Transcript: SOW-based vocabulary testing for NCELP Y7–principles, design, creation

Length of Talk: 19 minutes

Presenter: Natalie Finlayson

# Slide 1 (0.28):

Hello, this is Natalie, one of the resource developers at NCELP. You might have seen my name on some resources if you teach German, and I’ve just recently started working on French as well.

Together with Emma Marsden I’ve been leading on the vocabulary test development side of things, and I’m going to talk you through the steps involved in that process in this presentation. When you’re ready to start, just click to move onto the next slide, and the voiceover should start automatically.

# Slide 2 (0.34):

So, just to outline of the structure of this presentation. I’ll begin by acknowledging everyone involved in the vocabulary testing process, before going on to briefly talk about our aims and objectives for testing, and explaining the theoretical principles on which our test design is based.

I’ll then present the seven questions types you can expect to find in the Year 7 spring test, before talking a bit about the automatic scoring process.

Finally, I’ll conclude with a summary of the ways in which the design you’re about to see matches our testing aims.

# Slide 3 (0.43):

I’ll begin by thanking everyone involved in the vocabulary testing process, which as you can see was a huge team effort involving just about everyone in the NCELP office!

Particular thanks must go to the test development leads Kirsten Somerville, Ivan Avaca and Inge Alferink, who designed the questions for each language. Also, our programming expert Giuila Bovolenta, who advised on the process for coding these into electronic format.

As mentioned, however, almost everyone involved with NCELP has lent a hand in some way, from proofreading tests to compiling paper versions to coding question types to advising on design – it really gives an idea of the scale of the project, and how much work as gone into it.

# Slide 4 (1.24):

I’m moving on now to discuss the objectives we set ourselves for the testing process.

The overall aim of the vocabulary test is twofold. We want to find out not only *how many* of the words in the Year 7 schemes of work students have learned, which I’ll refer to as *breadth of knowledge* testing, but also *how well* students know these words, which I’ll call *depth of knowledge* testing.

To do this, we need to create a syllabus-based achievement test, which means we are only testing students’ knowledge of words encountered in the schemes of work so far. We are not testing their ability to work with cognates, or to infer meaning from context at this stage.

We bore three main considerations in mind when choosing question types. First, we wanted to make sure that students could not only *recognise* forms and meanings of L2 words when presented with them, but that they could *recall these* from their memories without the need for prompts. This consideration informed our decisions regarding question formats, such as the number of multi-choice and free production questions to include.

We also wanted to test students’ *receptive* and *productive* knowledge in equal measure, which meant that 50% of the questions needed to test listening and reading skills, and 50% written and spoken production. This, in turn, assured that students’ knowledge of both oral and written language was adequately assessed.

# Slide 5 (1.21):

To get an overall estimate of students’ breadth of knowledge, we would ideally like to test them on each word encountered in the schemes of work so far. Given that students will have learned over 200 words by this point in the year, however, this was deemed unfeasible in terms of both timing and concentration levels.

As the target length of the full Year 7 spring test is 40 minutes, divided between the strands of phonics, vocabulary and grammar, we estimated that a total test time of 12 minutes, or 720 seconds, was an appropriate aim for the vocabulary component. Pilot testing indicated that students could comfortably answer one question in 9 seconds, and so 80 questions in 12 minutes, giving us a nice round target figure of 20 questions for each of the modalities (listening, reading, writing and speaking).

As you can see, these figures results in high scheme of work coverage – 35% of words in the German and Spanish schemes of work will be tested, and 43% of those in French. This suggests the test results will give a highly reliable indication of students’ individual and collective performance – especially when you consider that some major online tests, such as the Vocabulary Size Test for learners of English, consider a coverage of just 1-4% sufficient for this purpose.

# Slide 6 (1.10):

Although we’ve established that 80 words is a very reliable sample size, it is still useful for us to test knowledge of all words in the schemes of work, so that we can identify any words which may pose more challenges than others, for example.

One way to do this is to distribute the words to be tested amongst the students. Qualtrics, which is the program we are using to host the tests, allows us to code questions for each vocabulary item in the scheme of work, and then present students with random samples of this. This has the added practical advantage that students cannot copy one another, as no two sets of questions will be exactly the same.

To ensure that students are presented with 80 *different* words (and not the same word tested in multiple ways) each word must be included in the test only once, meaning that 25% of words in the schemes of work are assigned to reading tests, 25% to listening tests, and so on.

The word class ratios of individual languages, shown here at the bottom of the slide, should also be broadly upheld within different question types, ensuring that students aren’t presented with a random sample containing only nouns, for example.

# Slide 7 (1.34):

I’m moving on now to discuss depth of knowledge testing. As previously mentioned, we want our tests to give us an idea of not only *how many* words students have learned, but *how much* they know about those words. So, we need to incorporate depth of knowledge testing into our design by choosing question types which assess different elements of *word knowledge.*

The parts of word knowledge on which NCELP vocabulary teaching is based draw on the aspects identified here by Paul Nation, who divides word knowledge into three main areas – form, meaning and use. If you’ve been adopting the NCELP schemes of work, you’ll notice that students’ first encounters with vocabulary items are in isolation, allowing students to focus on developing their knowledge of the spoken and written *form.* In future encounters with these words, which are incorporated systematically as part of our regular revisiting cycle, words are presented in context, to allow students to deepen their knowledge of *meaning* and *use.*

Not all elements of word knowledge will be tested, as not all have yet been covered in the schemes of work – students’ word pools simply aren’t big enough to incorporate teaching of word family structure or appropriate register use.

Knowledge of the grammatical patterns into which words enter is, of course, assessed in a separate part of the test.

The elements of word knowledge that remain – spoken and written form, all elements of meaning and collocation – are the elements on which our Year 7 teaching, and, therefore testing, is based.

# Slide 8 (0.20):

If you’re adopting our schemes of work, you may have come across this worksheet, created by Victoria Hobson and Rachel Hawkes, at the beginning or end of this half year.

It’s a checklist summarising of the elements of word knowledge we saw on the previous slide in student-friendly language, and we encourage its use for regular vocabulary revision.

# Slide 9 (0.46):

One final set of terms I’d like to define before moving on to present the question type themselves are *recall* and *recognition.*

Recognition tests usually take the form of multiple-choice activities in which students select or even guess the correct response from a list of distractors and exercises like this are designed to strengthen any existing memory traces of the words that the students already have.

Recall tests, on the other hand, do not include multi-choice options and demand the production of responses from memory so recall tests are often considered more difficult that recognition tests because learners have to actually reach into their memories and pull out those definition or those forms on their own.

# Slide 10 (0.55):

So, moving on now to present the seven question types we’ve identified as suitable for the Year 7 spring test.

On this and the slides that follow, you will see a screenshot of how questions are presented to students on the left, and a summary of the question type and knowledge tested on the right. Students who take the test will be presented with 10 examples of each question type 1-6.

Question Type 1, shown here, is a listening test which requires students to hear L2 words in isolation, and recognise their meanings from a set of pictures or English translations presented to them. Students hear words twice on clicking play, with a 1 second gap in between.

I’ve included a reference to Read (2000) in the top right here if you’re interested in reading more about the usefulness of out-of-context presentation of vocabulary for memorising meaning in the early stages of language learning.

# Slide 11 (0.38):

The second question is also a listening exercise, in which students must this time retrieve the meaning of an L2 word they have heard from memory, before assigning it to one of four English categories provided.

This is a meaning recall test which doubles as a ‘scaffolded’ version of a semantic relations exercise. In future years, category headings will also be given in the L2, and the exercise will become more about students showing that they can organise words into L2 hierarchies.

For now, this test requires students to demonstrate that they have understood the broader meanings of the words they have learned, in addition to the ability to produce rote-learned definitions.

# Slide 12 (0.30):

The third question type is a reading test in which students are asked to recall the English translations of L2 words presented to them.

This time, words are presented in a non-informative context. This shows students the *function* of the word, so there can be no doubt as to the word class the exercise intends to elicit, but gives very limited contextual clues so as not to give away the meaning.

Again, you can read more about this word presentation technique in Read (2000).

# Slide 13 (0.30):

The fourth question is probably the most challenging of the question types, as it requires students to read a prompt *and* a set of distractors in the L2, and recall the meaning of the words in each.

Students are then asked to select *two* words which could fill the gaps to make sensible L2 sentences. This scaffolded collocation exercise is very much in keeping with the NCELP principle of de-chunking, in that students are encouraged to demonstrate their ability to build novel sentences, rather than rely on rote-learned constructions.

# Slide 14 (0.46):

Question 5 is the first of two written exercises which ask students to recall the L2 form of an English word, and provide a translation. Words are presented with an informative context in both languages so that students understand the intended sense of the word, which they may have learned with multiple meanings.

It is important to note that students are not expected to draw on any grammatical knowledge when producing the words. The questions are designed so that expected answers will always take the forms that students have practiced on Quizlet. Nouns beginning with consonants will appear alongside their definite articles, nouns beginning with vowels alongside their indefinite articles, verbs in their infinitive forms, and adjectives in their masculine or feminine forms as indicated in the question.

# Slide 15 (0.43):

The second written exercise is also a form recall test, this time with the English context removed. Students are asked to complete an L2 sentence by translating an English word provided in brackets.

This is also a scaffolded collocation test. In future years, students will be required to determine the missing word from L2 context only, with the first letter given as a hint. In the second sentence on this slide, for example, the letter ‘r’ would appear at the start of the gap to direct students to the missing word *regarder*.

This type of test has been successfully carried out on a large scale by Laufer and Nation – again, see the reference in the top right for more detail.

# Slide 16 (0.41):

The final question type tests students’ ability to recall the spoken form of a given English word. Concrete words were selected for this question type to eliminate the need for context and, exceptionally, all 20 words in the spoken part of the test are tested in the same way.

This is the only question type that cannot be scored automatically. This question will not, therefore, appear in the main vocabulary test, but as part of a separate speaking test which also contains phonics and grammar elements. Students record themselves using *Vocaroo* software. This process is explained in Robert Woore’s [presentation](https://resources.ncelp.org/concern/resources/1j92g768v?locale=en) on phonics testing, found on the NCELP portal.

# Slide 17 (0.59):

In terms of question distribution, you’ll see on this summary slide that there is a strong bias toward recall in our test. There are two main reasons for this.

First, in keeping with NCELP principles, we want to encourage students to build a strong active vocabulary from the very beginning, and so we are testing them mainly on their ability to retrieve meaning and form from memory.

Second, a focus on recall gives us the freedom to include varied question types, and limit the number of multiple choice exercises. Overreliance on multiple choice allows students to gain a significant number of marks from guessing. In a 6-choice format, students can achieve a score of 16.7% without any knowledge at all. In a 4-choice format, this figure increases to 25%.

We do acknowledge that this results in a more challenging test, and we may revise the ratio in future years following analysis of students’ performance in this pilot study.

# Slide 18 (1.20):

The planning process for assigning words to question types involves the use of a very sizeable spreadsheet! Here, you see a snapshot of the planning phase for German.

One factor which must be considered in question allocation is the degree to which a word represents a concrete concept. More concrete items can be assigned to picture-matching tasks, while more abstract words lend themselves better to context-based exercises. The degree to which words are polysemous also plays a role, as this also affects the amount of context needed to ensure that the target meaning we wish to elicit is clear to students.

Thought must also go into choosing appropriate distractors for multiple choice questions. Tallies must be kept to record the number of times each word in the scheme of work appears as a distractor, so as not to over- or underrepresent any particular word.

Two main factors influence the choice of distractors which appear alongside a target item. First, all distractors must belong to the same word class as the target item, so that students cannot use their grammatical knowledge to select answers by process of elimination. Second, the spelling of words used as distractors plays a role in question differentiation – distractors similar in form to the target item, containing many of the same letters, add an extra challenge.

# Slide 19 (0.29):

Now let’s look at what is perhaps the biggest advantage of online testing for teachers – the autoscoring process.

Over the course of the next few slides, I will explain the steps involved in teaching the software to recognise correct answers, and the considerations involved in this.

For the multiple-choice questions which you can see on this slide, the process is fairly straightforward – as you can see here, target responses are simply assigned one point, and distractors zero points.

# Slide 20 (0.57):

For free production exercises, the process was a little more complex as we wanted to build in some tolerance for partially correct answers.

At this point in the Year 7 syllabus, students have had only 18 weeks of lessons, and we want to be fully tolerant of accent errors at this stage. As you can see on the slide, this means that responses which contain a missing or incorrect type of accent, but are otherwise correctly spelled, are awarded one point.

We also want to award partial credit for correctly spelled words with missing or incorrect articles. This may have limitations in the weighting that this gives to knowledge about grammatical gender, and the ratio may be given more consideration in future iterations of the online test.

This year, however, we need to avoid the need for overly complicated mathematics as we are also creating paper versions of the test, which will be marked manually. So, we’ve decided on a 50:50 split this year for consistency.

# Slide 21 (0.23):

In some cases where the target word requires both an article *and* an accent, this manual coding process is somewhat lengthy, as you can see in this quite fitting example (!) of the French word *le problème.*

The big advantage, of course, is that this process is a one-off time investment - the encoding of each item need only be completed once, and can then be reused year after year.

# Slide 22 (0.31):

Even more challenging to program was a level of tolerance which took into account English misspellings in the L2 to L1 translation task.

We have already encoded some of the most frequent misspellings based on our intuition, as you can see in the example of *friend* here. However, as it is impossible to predict all the creative misspellings of a word that students might come up with, this is the one section that does need to be manually checked, at least in the beginning as we build up a ‘bank’ of acceptable misspellings in the program.

# Slide 23 (0.53):

I’ll now conclude with a recap of the key features of the NCELP vocabulary test, summarising the links between the testing objectives we identified, and the design decisions we made.

The NCELP Year 7 vocabulary test is a syllabus-based vocabulary test designed to measure vocabulary breadth and depth. It provides a highly reliable snapshot of student achievement in a manageable timeframe and tests all words featured in the scheme of work by distributing them randomly amongst students. It assesses recognition and recall skills across the four modalities (reading, listening, speaking and writing). It thoroughly tests different element of world knowledge in line with aspects taught in the scheme of work. It provides automated scoring for 6 out of 7 question types.

All in all, I would say this is quite good going for a 12 minute test!

# Slide 24 (0.31):

That’s the end of the presentation. I hope it has been useful! You’ll find references to the research I have referred to in the presentation in the bibliography on the next slide.

If you have any questions at all about the theory on which we have based the test design, or would like to talk about ways to adapt the design to suit your own syllabus, just send me an email at Natalie.Finlayson@york.ac.uk, or tweet me @natalie\_eloise.

Thanks a lot for listening, and goodbye!

# Slide 25:

No audio.