
Article:
Finlayson, Natalie Eloise, Marsden, Emma orcid.org/0000-0003-4086-5765 and Anthony, Laurence (2023) Introducing MultilingProfiler: An adaptable tool for analysing the vocabulary in French, German, and Spanish texts. System. 103122. ISSN 0346-251X

https://doi.org/10.1016/j.system.2023.103122

System 118 (2023) 103122

Contents lists available at ScienceDirect

System

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

Introducing *MultilingProfiler*: An adaptable tool for analysing the vocabulary in French, German, and Spanish texts

Natalie Finlayson [a,*], Emma Marsden [a], Laurence Anthony [b]

[a] Department of Education, University of York, York, YO10 5DD, UK
[b] Center for English Language Education (CELESE), Faculty of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555, Japan

ARTICLE INFO

Keywords:
Vocabulary
Lexical profiling
Materials development
Language testing
Methods
French
German
Spanish
Language education

ABSTRACT

Vocabulary profiling with computational tools and word lists is an established step in the development of pedagogical materials for learners of English. However, existing tools and word lists lack sensitivity to the orthographical, morphological, and grammatical systems of highly-inflected and declined languages. This limits the degree to which lexical profiling can be usefully implemented in the creation of materials intended for use with beginner/low-intermediate learners of such languages who have only partial knowledge of these systems.

In this article, we present *MultilingProfiler*, a vocabulary profiling tool designed to support nuanced profiling of texts in French, German, and Spanish. We introduce the concept of 'bespoke' word families tailored to the needs of learners at various stages of development, and outline key features of the tool that operationalise this concept (the functionality to select which inflected, derived, and multiword forms of headwords are included in the profile; sensitivity to orthographical systems; embedded word lists aligning with specific programs of study; and cumulative word lists that grow with learner knowledge). We present two case studies that find *MultilingProfiler*'s features to be effective in highlighting potential mismatches between the lexical demands of texts and the expected knowledge of learners, and consider applications of the tool in research methods.

1. Introduction: Lexical profiling and language learning

In its broadest sense, lexical profiling is the practice of using software to measure the distribution of lexical items in texts across one or more word lists. Such analysis can provide an indication of the lexical load of a text, the potential for acquiring or developing vocabulary knowledge by reading or listening to a text, and the nature of vocabulary knowledge needed to understand different spoken and written text types (Nurmukhamedov & Webb, 2019). In the pedagogical context, this information allows texts to be classified by vocabulary level, and selected or adapted for use with a particular learner group accordingly (Dang, 2023).

Lexical profiling is an established step in the development of educational materials (learning resources, programs of study, and tests) that support vocabulary learning and teaching in English as a Second Language (ESL) and English for Specific Purposes (ESP) curricula. Its application at different stages in learning for a range of purposes has been made possible by an extensive body of research that informs the development of English word lists (see section 2.1) and the emergence of a number of freely available tools that

* Corresponding author.
E-mail addresses: Natalie.Finlayson@york.ac.uk (N. Finlayson), Emma.Marsden@york.ac.uk (E. Marsden), anthony@waseda.jp (L. Anthony).

https://doi.org/10.1016/j.system.2023.103122
Received 20 July 2023; Received in revised form 13 August 2023; Accepted 17 August 2023
Available online 30 September 2023

support the use of these word lists in textual analysis (see section 2.2). Some such tools can also be used with a non-English language word list (e.g., *VocabProfilers*, Cobb, n.d.; see section 2.2), to create (f)lemma or word family-based profiles which are useful in the development of materials for intermediate and advanced language users. However, a lack of sensitivity to the orthographical, morphological, and grammatical systems of highly-inflected and declined languages limits the effectiveness with which such tools can be used to create profiles that represent the partial vocabulary knowledge of beginner and low-intermediate learners. As a result, the lexical profiling step is likely to be less widely implemented in the development of materials for learning and teaching inflected and declined languages than it is for English.

MultilingProfiler (Finlayson, Marsden, & Anthony, 2022) is an online, freely available (under a CC BY-NC-SA 4.0 licence) lexical profiling tool optimised for use with word lists and texts in French, German, and Spanish. Its key features—sensitivity to French, German, and Spanish orthography; the functionality to select which inflected, derived, and multiword forms of headwords are included in the profile; the embedding of bespoke word lists that align with learners' knowledge about language at specific points in a program of study; and the capacity to profile against a combination of user-made lists and built-in lists—address some of the limitations of existing profiling tools in the analysis of texts intended for use with beginner and low-intermediate learners.

A practical, context-specific incentive to develop *MultilingProfiler* was provided by language education policy changes in England (Department for Education, 2022) that require Awarding Organisations (developers of high stakes national examinations) to provide lists of the words that can be tested in the GCSE[1] examinations taken by students of French, German, and Spanish at age 16. The changes were driven by a desire to improve the uptake and quality of the study of these languages in secondary schools (e.g., Teaching Schools Council, 2016), and word lists were introduced with a view to helping educators and materials developers better align the vocabulary taught and tested in schools. Thus, many aspects of *MultilingProfiler*'s features were designed with the pedagogic needs of beginner or low-intermediate learners with English as a first language (L1) in mind. However, we also wanted to develop a vocabulary profiler that could support the selection and creation of levelled texts in French, German, and Spanish more broadly. To this end, we developed *MultilingProfiler* in line with the following aims: (i) to develop a vocabulary profiling tool that supports the flexible profiling of texts in French, German, and Spanish, and (ii) to develop an approach to bespoke word list creation that supports the profiling of texts in French, German, and Spanish for particular learner groups.

In this article, we discuss theoretical considerations relevant to lexical profiling in French, German, and Spanish, and the extent to which existing word lists and tools are equipped to deal with the challenges posed by highly-inflected and declined language systems. We then present features of *MultilingProfiler* that are designed to address limitations of existing word lists and tools in this regard, and two illustrative case studies that showcase pedagogical applications of these features. Finally, we consider future applications of the tool in language learning research methods.

2. Literature review

With the context of learners of French, German, and Spanish at different proficiency levels in mind, we first review literature on the purpose and development of pedagogical word lists (in section 2.1) before examining features of existing vocabulary profiling tools to identify gaps in provisions (in section 2.2).

2.1. Word lists for learners of French, German, and Spanish at different proficiency levels

We have identified two bodies of literature key in informing the development of lexical profiling tools intended for use in the design of pedagogical materials: those that report on the importance of lexical coverage (readability) measures for different learning activities (section 2.1.1) and those that describe the challenges of defining an appropriate unit of word counting for different proficiency levels in different languages (section 2.1.2).

2.1.1. The lexical coverage construct

The most basic function of a lexical profiling tool is to provide a measure of lexical coverage, that is, a calculation of the percentage of words in a text that appear on a given word list. Where a pedagogical word list is used, this figure should represent the total coverage of a text by words that a learner is considered or expected to understand at a certain stage in their learning. Depending on the aim of the learning activity, it may be desirable for this measure to be high, as in the case of meaning-focused activities and fluency development activities (Dang, 2019; Nation, 2022), or lower, if inferential or dictionary skills are the focus. The most fundamental application of lexical profiling software in materials design, then, is to assist with the process of selecting, adapting, or creating texts to achieve a lexical load deemed appropriate for the purposes of a specific task. For intermediate and advanced learners of English, the lexical coverage necessary for adequate comprehension of texts has been shown to be somewhere in the range of 95–98% for written materials (Hu & Nation, 2000; Laufer & Ravenhorst–Kalovski, 2010) and 90–95% for spoken materials (Giordano, 2021; van Zeeland & Schmitt, 2013). If the acquisition of new words from extensive reading is expected, the figure should be 98% (Schmitt, Jiang, & Grabe, 2011). If minimal comprehension with some inferencing of unknown items is expected, 90% coverage may suffice (Laufer, 2020).

Few studies have been carried out to investigate the lexical coverage needed for the successful completion of different learning

[1] The General Certificate of Secondary Education (GCSE) is a high stakes external examination taken by students in England and Wales at the end of Year 11 (ages 15–16). Performance on components of the GCSE have been estimated to represent language proficiency equivalent to A1-A2 on the Common European Framework of Reference (Curcin & Black, 2019).

tasks in French, German, and Spanish, though we are aware of two exceptions. Herman & Leeser, 2022 tested the 95–98% coverage range for reading with beginner learners of Spanish. They found 98% coverage to be optimal, though noted variation in the degree of comprehension of longer texts by learners who nevertheless seemed to have similar word knowledge. Noreillie et al. (2018) investigated the coverage required for oral comprehension of intermediate texts in French, and reported results that align with van Zeeland & Schmitt (2013) (i.e., 90–95%). The stipulations for alignment between word lists and exam texts in the new GCSE subject content (Department for Education, 2022) draw on these findings, though err on the side of providing additional supportive constraints. With the exception of proper nouns, 100% of L2 words in that appear in listening transcripts and 96% of L2 words that appear in written text must appear on the prescribed word lists, with the 4% of off-list words in reading texts comprising glossed words and cognates only.

The range of appropriate coverage measures for different levels, languages, and purposes points to the importance of the vocabulary profiling stage in the production of materials intended to align with the perceived needs of individual groups. Specifically, a lexical profile can alert users to (i) the potential lexical demands of a given text and, thus, its appropriacy for the intended task; (ii) any vocabulary items that may need additional attention; and (iii) the potential to learn unfamiliar words by reading or listening to a text.

2.1.2. Defining and counting words at different stages of learning: The need for a flexible profiler

Most English word lists used with existing lexical profiling software count lemmas (e.g., Brezina & Gablasova, 2017), flemmas (e.g., Browne, 2014), or word families at level 6 (WF6) on Bauer and Nation's (1993) scale of affix frequency, productivity, predictability, and regularity (e.g., Coxhead, 2000; Nation, 2017a; West, 1953). Lists that count lemmas consider a 'word' to consist of a headword and the inflected forms of that headword (e.g., the count for *focus* [verb] includes *focus, focuses, focussing,* and *focussed*; the count for *focus* [noun] includes *focus* and *foci*). Flemma-based lists count similarly, but do not distinguish between parts of speech (e.g., the count for *focus* [noun or verb] comprises *focus, foci, focuses, focussing,* and *focussed*). Lists based on WF6 include the headword, derived forms of that headword, and the inflected forms of the headword and each of its derived forms (e.g., the count for the base word *focus* includes *(re)focus, foci, (re)focuses, (re)focussing, (re)focussed*, and *unfocussed*). Use of types (individual word forms, i.e., considering *focus, focuses, focussing, focussed*, and all related derivations as separate units) has been largely discounted as a viable option for pedagogical word lists, given the 'very pessimistic assumption' (Bauer & Nation, 1993: 258) that learners of English could ever be at a level where they are unable to deal with variation in written inflectional form.

As the purpose of lexical profiling in the pedagogical context is to assess the suitability of a text for a target group of learners, the unit of counting chosen must relate to the learning required or expected (Nation, 2016). We are of the view that none of the units described above—type, lemma, flemma, or WF6—is likely to be a perfect fit for beginner and low-intermediate learners of French, German, and Spanish. While it might be reasonable to assume that learners of English can recognise relatively small sets of inflected forms (e.g., *book* and *books*) as members of the same family once they have mastered the headword (Bauer & Nation, 1993), there is evidence to suggest that counting lemmas or flemmas may be too great a step in the initial stages of acquisition of highly-inflected and declined languages (for an illustration of the differences in these systems, see Tables A.1 and A.2 in Appendix A). It is known that highly irregular inflections of English words (e.g., *went*) are stored in the mental lexicon as distinct holistic forms (see e.g., Kempley & Morton, 1982; Marslen-Wilson, Hare, & Older, 1995; Pinker, 1991; and also Meunier & Marslen–Wilson, 2004, for a detailed overview of studies in English). The same applies to French verb forms with irregular stems (e.g., the imperfect *buv*-stem of the verb *boire* ['to drink']) (Estivalet & Meunier, 2015). The implications of this are limited to the learning and teaching of a relatively small number of irregularly inflected words in English, but the issue is likely to be amplified in French, German, and Spanish. For example, it seems unlikely that a beginner/low-intermediate learner of French would intuitively recognise that the frequent but irregularly inflected headword *être* ('to be'), its present tense forms (*suis*, *es*, *est*, *sommes*, *sont*), and its imperfect (*ét-*), past historic (*fu/fû-*), subjunctive (*soi-/soy-*), and future/conditional (*ser[i]-*) stems are part of the same lemma without receiving direct instruction of, or massive exposure to, these paradigms (though this has not been empirically tested). The same issue applies to the recognition of some less frequent members of regularly inflected lemma sets, for example, the subjunctive *toquemos* ('let us play/touch') and dative plural *Büchern* ('books') forms of headwords *tocar* ('to play') and *Buch* ('book') in Spanish and German. In the case of these regularly inflected forms, challenges may arise from the significant decrease in the proportion of the word represented by the known stem, as in the addition of an umlaut and three-letter suffix to *Buch* in the dative plural form (see e.g., Dijkstra et al., 2010, on the role of orthographic overlap on word recognition, albeit cross-linguistically with cognates). This suggests that some holistic representation of particular inflected forms (types) could be usefully included as individual entries on a beginner or low-intermediate word list.

Moreover, the complexities of French, German, and Spanish lemmas described above challenge the belief that the lemma (approximately equivalent to word family level 2 in Bauer & Nation's [1993] taxonomy) is a more appropriate unit of counting for evaluating the vocabulary in texts intended for beginner and low-intermediate learners than a word definition which includes derived forms. This notion may well hold for English, with a number of studies suggesting that advanced learners of English are more likely to recognise WF6 derivatives or known words than learners at lower levels (e.g., Brown et al., 2020; Laufer et al., 2021; Snoder & Laufer, 2022) and others yielding mixed results (e.g., McLean, 2018; Ward & Chuenjundaeng, 2009). We are not aware of any studies that have looked at this issue in the context of French, German, and Spanish. However, it seems plausible that frequent, productive, predictable, and regularly derived forms of known headwords in these languages (word family level 3 [WF3] in Bauer and Nation's terms) may be more straightforward to recognise than some irregularly inflected forms, particularly in our UK context where learners with knowledge of English may be able to draw on similarities between closely-related affixation systems (Laufer, 2021). Potential examples include the addition of *in-*to French adjectives, adverbs, and nouns where the English equivalent is 'un-' or 'in-'; the addition of *-los* to German nouns to create adjectives where the English equivalent is '-less'; and the addition of *-able* to Spanish verb stems where the English equivalent is also '-able'. This supposition also depends to some degree on the amount of exposure learners have had to the derivational affixes concerned, a topic which, Dang (2022) notes, has received very little attention in language education textbooks compared with instruction of inflected forms. In response to this

general trend, the new GCSE content (Department for Education, 2022) specifies a set of frequent, regular, predictable, and productive affixes that may be assessed in reading. New instructional material for teaching these patterns (e.g., NCELP, 2022a), may provide opportunities for further research with learners who have undertaken practice to systematically acquire knowledge of derivational morphology in coming years. The results of such studies could yield important insights into the ability of these learners to recognise parts of the inflectional and derivational systems of French, German, and Spanish, which may well differ from the findings of studies carried out with learners of English whose knowledge was acquired implicitly (Iwaizumi & Webb, 2022).

In sum, although the paucity of research in our target languages with different kinds of learners means that our suppositions remain tentative, current evidence suggests that lemmas may be too broad a unit of counting at beginner and low-intermediate levels, whereas some derivational morphology may be appropriately incorporated into the word definition. With this in mind, we take inspiration from those who emphasise the importance of varying the selection of lexical units according to pedagogical purpose and learner variables (see e.g., Bauer & Nation, 1993; Nation, 2017a, 2017b; Cobb & Laufer, 2021; Webb, 2021a, 2021b), and argue the need for a profiling tool that can tailor the unit of counting to the lexical and grammatical knowledge of the learner group, be this unit at the level of type, lemma, flemma, word family, or any combination. Henceforth, we use the term 'bespoke word family' to describe such units of counting, which may incorporate types, parts of lemmas, and/or parts of word families as appropriate.

2.2. Lexical profiling tools

As mentioned in the introduction, a wealth of vocabulary profiling tools for use with different word lists is available to ESL and ESP practitioners. *AntWordProfiler* (Anthony, 2022a) is a modern desktop profiling tool that extends the functionality of its predecessor *Range* (Heatley et al., 2002). It supports the profiling of texts with two embedded lists—the General Service List (West, 1953) and the Academic Word List (Coxhead, 2000)—and any other word list imported by the user. Links to two further widely used word lists—the New General Service List (Browne, 2014) and the BNC/COCA lists (Nation, 2017a; 2017b)—are provided on the *AntWordProfiler* homepage. *VocabProfilers* (Cobb, n.d.) is a web-based alternative that offers profiling with embedded versions of all four aforementioned lists, and also the Nuclear Word Family List (Cobb & Laufer, 2021), the BNC-COCA Common Core List (Gardner, 2013), and the Billuroğlu–Neufeld List (Billuroğlu & Neufeld, 2007) which is an amalgamation and extension of other lists mentioned here. *LancsLex* (Brezina & Gablasova, 2017) is another online tool that offers profiling with the New General Service List (Brezina & Gablasova, 2015) and the 10,000 most frequent words in the British National Corpus 2014 (Brezina et al., 2021; Love et al., 2017). Other web-based tools support profiling with English word lists that are considered to align with different levels of the Common European Framework of Reference for Languages (CEFR) as well as other general and specialised word lists. These include *Text Inspector* (Bax, 2012), *VocabKitchen* (Garner, 2022), the *New Word Level Checker* (Mizumoto, 2021), and the EAP Foundation's *Vocabulary Profiler* (Smith, 2022).

We are only aware of three tools that support the lexical profiling of texts in one or more of French, German, and Spanish. These are *VocabProfilers* (Cobb, n.d.), *AntWordProfiler* (Anthony, 2022a), and *Sketch Engine* (Kilgarriff et al., 2014). *VocabProfilers* offers frequency-based profiling in French with embedded lists informed by the rankings in Lonsdale & Le Bras (2009). The interface allows users to enter texts of up to around 25,000 words and perform lexical coverage calculations using any frequency band (in 1,000-word increments) up to the 25,000 level. It offers a degree of customisation in that it allows users to specify sets of words (e.g., cognates, proper nouns, compounds) to add to the lists and include in the profile. Statistics about the number, nature, and frequency bands of the words in the text are provided at the levels of lemma, type, and token. The tool also allows users to carry out a cognate analysis of texts in French by comparing the words they contain to English items appearing on the BNC-COCA frequency lists. *AntWordProfiler* does not come with word lists for French, German, and Spanish built in, but can be used with word lists in any language sourced and uploaded by the user. It does not support the addition of items to imported word lists, but does allows users to specify a set of words to eliminate from the profile completely. For individual text profiling, *AntWordProfiler* uses a colour-coding system to indicate the lists in which the words appear. In this display, the tool provides information about the number and nature of types and tokens in the text. Both *VocabProfilers* and *AntWordProfiler* display the results of a variety of lexical measures, including type, token, and headword counts, in a set of statistical tables. *Sketch Engine* supports lemmatisation and part of speech tagging in 39 languages (see Appendix B), and its 'Wordlist' feature can be used with texts or corpora of texts to count the number of instances of types, tokens or (f)lemmas that appear on any word list added by the user. However, unlike *VocabProfilers* and *AntWordProfiler*, *Sketch Engine* does not provide embedded (fixed) word lists or a profiling interface for visualisation and text editing purposes. Instead, users must upload the texts they wish to profile as corpus files. *Sketch Engine* users must also pay a monthly subscription fee after a 30-day trial.

All three existing tools that are compatible with non-English word lists and texts require users to choose tokens, types, or (f)lemmas as the counting unit (in the case of *VocabProfilers*, a word families option is in development at the time of writing). As commonly-used tagsets for French, German, and Spanish include irregular and infrequent inflected forms of headwords in their (f)lemma sets, the outputs of profiles created using (f)lemmatised word lists are not sensitive to the issues of word counting relating to beginner/low-intermediate learners that we have discussed. Another limitation of these tools is that they process profiling output following the morphological rules of English. Thus, they are unable to recognise, for example, French words with hyphens (e.g., *peut-être* ['maybe']), apostrophes (e.g., *aujourd'hui* ['today']), or blank spaces (e.g., *parce que* ['because']) as part of the spelling of single words, treating them instead as multiword units with multiple components.

3. Features of *MultilingProfiler*

From our review of pedagogical word lists and existing vocabulary profiling tools, it is clear that resources to support the nuanced degree of profiling that may be necessary for pedagogically-oriented texts in French, German, and Spanish are lacking, particularly

where such texts are intended for use in the early stages of learning. Here, we introduce *MultilingProfiler* (Finlayson, Marsden, & Anthony, 2022), a free vocabulary profiling web app whose functionalities should offer some solutions to the specific issues we have identified. We briefly describe the *MultilingProfiler* interface (section 3.1) before returning to the two aims that informed its design: (i) to develop a vocabulary profiling tool that supports the flexible profiling of texts in French, German, and Spanish, and (ii) to develop an approach to bespoke word list creation that supports the profiling of texts in French, German, and Spanish for particular learner groups. In section 3.2, we address the first aim by presenting features of the tool that enable customisation of default counting units in existing (f)lemma-based frequency lists and deal with the orthographical systems of multiple languages. In section 3.3, we respond to the second aim by describing tool-independent approaches to creating word lists that count bespoke word families aligned with a particular program of study, taking as an example the context of the GCSE language curriculum in the UK.

3.1. MultilingProfiler interface

On accessing the website (https://www.multilingprofiler.net/), users are directed from the landing site to the 'MultilingProfiler' tab (see Fig. 1). Here, they are invited to type or paste a text into the profiling window (currently, texts of up to around 100,000 words can be processed in a few seconds) and select the list that they wish to use from the four available options in the 'List type' dropdown: frequency lists, lists aligned with the curricula of the language GCSEs, lists aligned with a particular stage in GCSE-level study, and custom lists provided by the user.

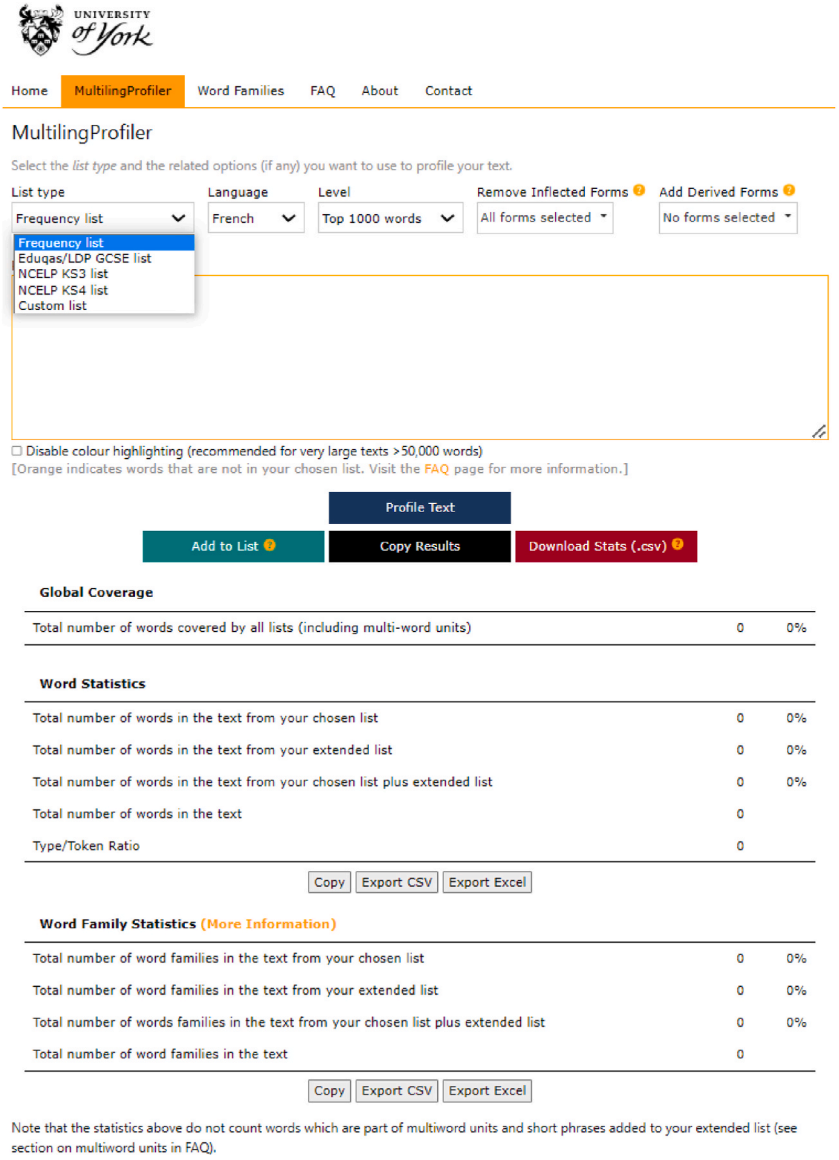

Fig. 1. *MultilingProfiler* interface.

MultilingProfiler then creates a visual profile using a colour-blind friendly palette (see Fig. 2). Any words in the text that do not appear on the selected list turn from black (indicating that they are 'on list') to orange ('off list'). Users can customise their chosen list by adding words to it using the 'Extended List' function. Clicking on 'Add to List' below the profile window opens a new window to the left where users can type any additional words or phrases that they wish to include in the profile. For example, users may consider proper nouns, cognates (as in the example in Fig. 2), or words that will be glossed to count towards a specific coverage figure, as these do not impede comprehension. Users can also group any words or phrases added this way into word families by listing them sequentially after a headword followed by a colon (e.g., adding 'USA: USA, USofA, US, United States, United States of America' to the extended list tells *MultilingProfiler* that these variations of the country name should be considered 'on list', and that they are part of a word family comprising five lexical items with headword 'USA').

MultilingProfiler uses the information in the selected list and extended list to provide a range of statistical data about the lexical composition of the text, and displays it directly below the profiling window. The most fundamental of these calculations is the 'Global Coverage' statistic, which is simply the total number of black (on-list) words in the profile and the lexical coverage these words provide. The 'Word Statistics' and 'Word Family Statistics' tables provide a breakdown of this information by displaying the number of, and coverage provided by, (i) lexical units in the text that are on the word list; (ii) lexical units in the text that are on the extended list, but not the word list; (iii) lexical units on the word list and the extended list combined. The total number of lexical units in the text is

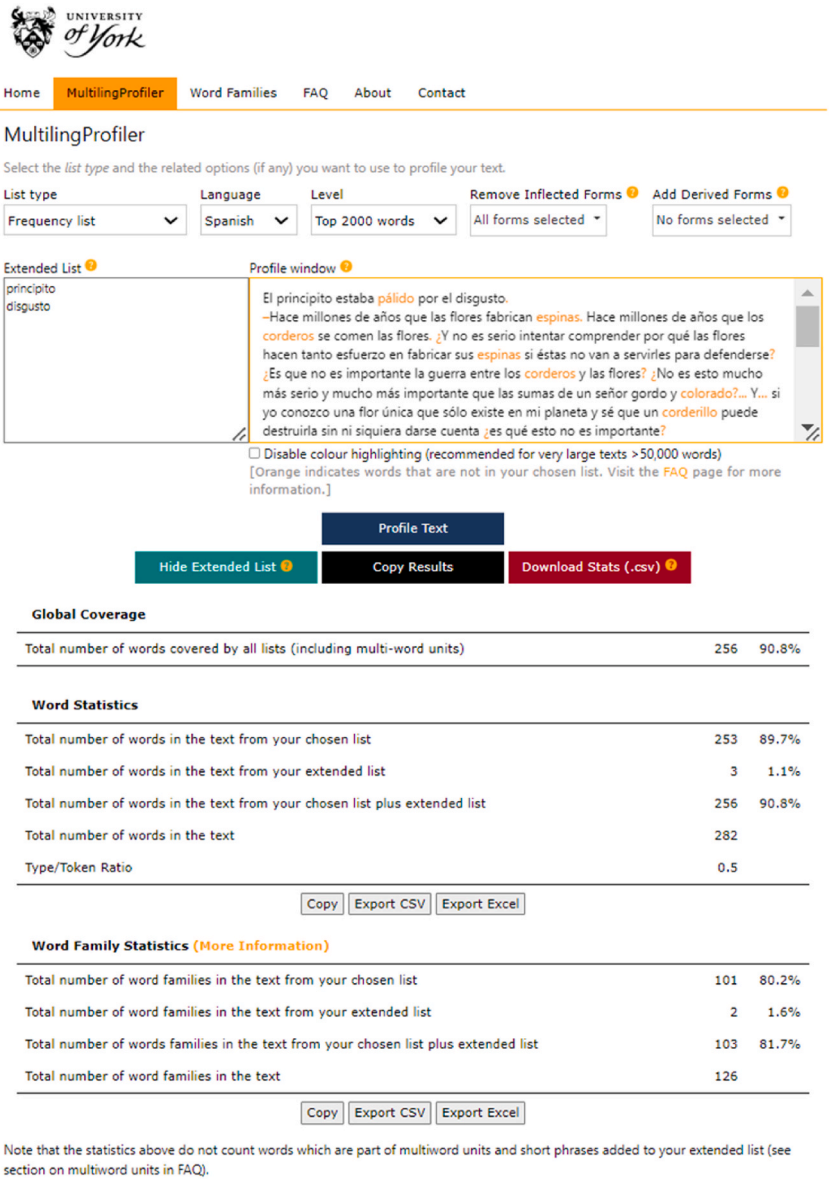

Fig. 2. Profile for *El Principito* against 2,000 band in Spanish.

also displayed. The 'Word Statistics' table counts individual words (tokens), while the 'Word Family Statistics' table counts word families according to the parameters specified by the user (see section 3.2) or word list (see section 3.3).

It is possible to edit texts directly in the profiling window and immediately reprofile them. No screen refresh is needed to see the colour highlighting and status updates.

3.2. *Creating bespoke word families: 'Remove inflected forms' and 'Add derived forms'*

The function of the 'Frequency list' list type is to create text profiles using embedded lists of the 1,000, 2,000, 3,000, 4,000 or 5,000 most frequently used words in each language. The lists are informed by the frequency order provided in the Routledge dictionaries of French (Lonsdale & Le Bras, 2009), German (Tschirner & Möhring, 2019), and Spanish (Davies & Davies, 2019), which are based on outputs from large general corpora that comprise balanced samples of written and spoken text. The Routledge dictionaries are, to the best of our knowledge, the most comprehensive and up-to-date general service lists for these languages currently available in the public domain (others may be added in future).

For the most part, the default unit of counting adopted for frequency-based profiling is that which is used in the Routledge source list, that is, flemmas for French, and lemmas for German and Spanish. For example, the French source list does not distinguish between the lexemes *son* (pron) ('his, its') and *son* (n) ('sound'), instead treating inflected pronominal forms (e.g., *sa* ['hers, its']) and nominal forms (e.g., *sons* ['sounds']) as parts of the same headword *son* (pron/n), ranked eighteenth in the frequency list. However, a distinction is made between the German lexemes *sein* (v) ('to be'), ranked fourth, and *sein* (pron) ('his, its'), ranked fortieth. Thus, any inflected forms of *sein* (v) (e.g., *bin* ['I am'], *sind* ['we/they/you are'] and *war* ['I/s/he was']) and declined forms of *sein* (pron) (e.g., *seine* ['his, its + feminine or plural noun in nominative'] and *seinem* ['his, its + masculine or neuter noun in dative']) in texts are considered parts of different words and counted separately. There is a caveat to this in that *MultilingProfiler* does not currently perform part-of-speech tagging of texts, and has not yet been trialled on its ability to work with texts that have been morphosyntactically tagged using other software such as *TagAnt* (Anthony, 2022b). Thus, the profiling output does not distinguish between the different parts of speech of polysemous and homonymous words. So, while the inflected and derived forms of *sein* (v) and *sein* (pron) can be counted separately because there is no orthographic overlap, *MultilingProfiler* cannot currently distinguish between the two meanings of *sein* itself, and addressing the limitations this introduces at the level of sense disambiguation is a work in progress (see section 5). To create the base word lists, we (f)lemmatised the headwords in each Routledge list using TreeTagger (Schmid, 1994), and passed the resulting lists of forms to language specialists and native speakers for manual checking. Outdated and erroneous word forms were removed, and a number of missing paradigms added.[2] The process of updating the (f)lemmatised lists through the addition of missing or new words as flagged by users is ongoing, and we hope that wider use of the tool will provide us with additional user-based feedback.

As we have discussed, however, (f)lemmas may not always be the most appropriate lexical unit for profiling French, German, and Spanish texts for beginner/low-intermediate learners because of the size and complexity of some morphological systems in these languages. *MultilingProfiler* accounts for this with two features that enable users to adapt the composition of the default (f)lemma to create a bespoke word family. The information about words included in the TreeTagger (Schmid, 1994) tagset (e.g., case, tense, part of speech) makes it possible to instruct *MultilingProfiler* to exclude certain word forms from the (f)lemma when users choose to deselect grammatical paradigms with the 'Remove Inflected Forms' dropdown, as in Fig. 3 (for the full list of grammar features that can be (de) selected for each language, see Table C.1 in Appendix C). Some checking of the outputs created using this feature is necessary in cases where word forms have multiple functions (e.g., the French *sois* ['I/you were, be']) is both a present subjunctive and imperative form, so will appear off list only if *both* 'present subjunctive' and 'imperative' are deselected).

Similarly, users can add some derivational morphology to the counting unit by selecting from the set of frequent, productive, predictable, and regularly derived affixes (as specified by Department for Education, 2022; see Table C.2 in Appendix C) in the 'Add Derived Forms' dropdown. To make the WF3 lists behind this feature, we first retrieved all theoretically possible derived forms of each word on each language's frequency list using the headword stem and a dictionary reference list, and then manually checked every entry returned to ensure relevance, accuracy, and completeness. Currently, the option to include derived forms is available for the 1,000 and 2,000 bands, with work on the 3,000–5,000 bands in progress.

Once users have defined the parameters of their bespoke word family, they can obtain detailed statistical data about their text by clicking on the 'Download Stats' button beneath the profile window (see Fig. 1). These data include a breakdown of the words in the text by frequency band; a breakdown of the words in the text by bespoke word family, including a raw count of each word type included in the family; and information about the number of occurrences of specified multiword units that appear in the text. By default, *MultilingProfiler* considers the words of which multiword units consist as single words that are 'on list' if they are included in the specified word list or extended list (i.e., they appear black in the profiling window). For example, the words in the phrase *faire des courses* ('to go shopping') are treated as parts of three separate word families: *faire* ('to make, do), *de* ('of'), and *course* ('errand'), and are included in the statistics for those word families. However, the tool also gives users the option to analyse multiword units as lexical units in their own right. Users can specify any multiword units and associated inflected forms that they wish to treat as a multiword family using the extended list function (e.g., adding *faire des courses*: *faire des courses*, *fais des courses* ['I/you go shopping'], *fait des*

[2] Prior to manual checking, missing paradigms in the TreeTagger (Schmid, 1994) lists were: (i) second person singular imperative forms of German verbs; (ii) inflected forms of German ordinal numbers; (iii) outdated orthographical use of the German Esszett (*ß*) rather than the trans-literate 'ss'; (iv) suffixed cliticised pronouns on imperative, infinitive, and gerund forms of Spanish verbs; (v) unaccented forms of French words normally beginning with accented letters in sentence-initial positions.

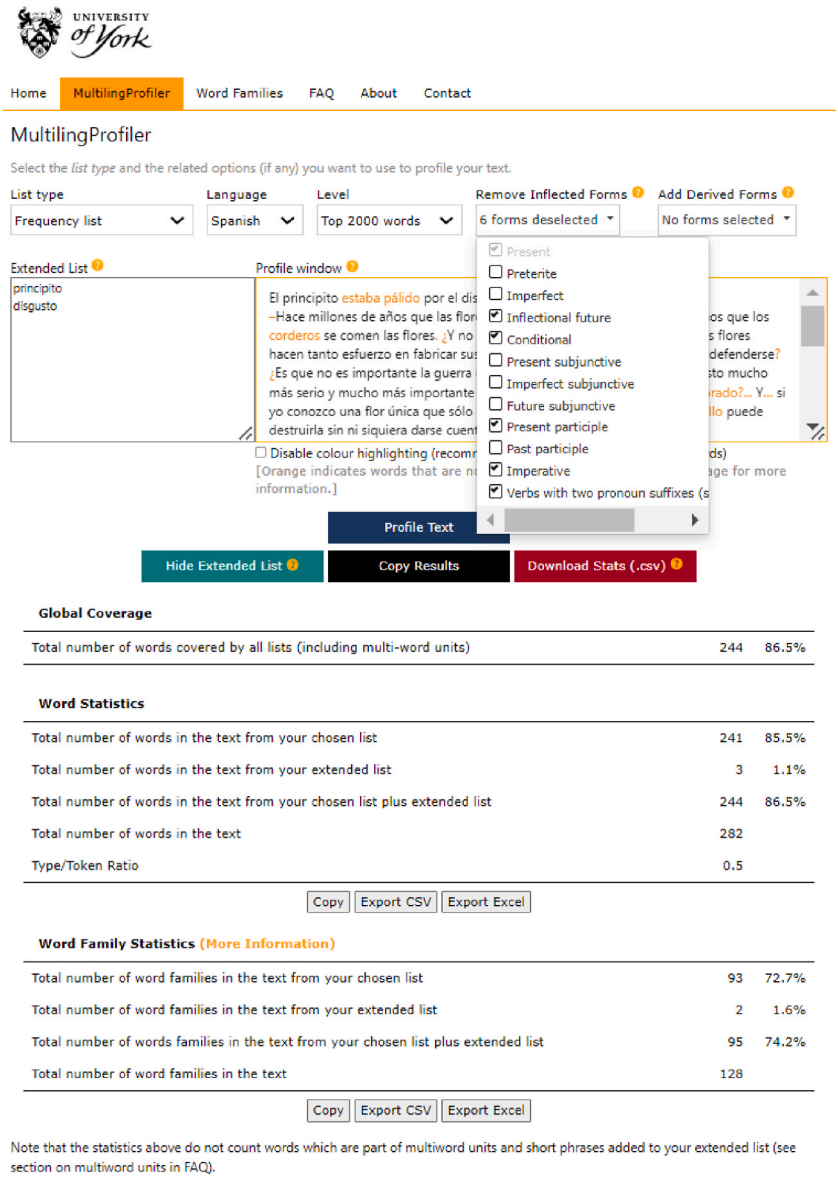

Fig. 3. Profile for *El Principito* against 2,000 band in Spanish with past tense forms and subjunctive forms deselected.

courses [‘s/he/it/one goes shopping’] creates a new word family of three lexical units with *faire des courses* as the headword). A multiword unit statistics file that counts occurrences of each multiword unit and multiword unit family can be obtained by clicking on ‘Download Stats’ (these advanced data are not reflected in the basic statistics tables on the main profiler tab).

To account for differences in the orthographical systems of the three languages, we had to consider the boundary markers used to define words from a spelling perspective. As discussed in section 2.2, a number of French words contain a hyphen as part of the orthography for a single word (e.g., *peut-être* [‘maybe’], *quatre-vingts* [‘eighty’], *lui-même* [‘himself’]). In German and Spanish, hyphens are used to connect two related (but separate) words, for example, the parts of a compound adjective (e.g., *blanco-azules* [‘bluish white’]). In German, hyphens also show items in a list that contain a common element (e.g., *Ein-und Zweibettzimmer* [‘one- and two-bed rooms’]). The boundary marker for French, therefore, differs from the other languages in that it considers a hyphen part of an allowable letter string (word). A workaround had to be implemented to ignore this rule in French questions, where inverted subjects and verb forms connected by a hyphen (e.g., *voulez-vous* [‘do you want’]) must count as two words rather than one. A similar workaround allows French words that contain apostrophes or blank spaces as part of the spelling (e.g., *aujourd’hui* [‘today’], *parce que* [‘because’]) to count as single words, ignoring the global word definition that counts letter strings containing apostrophes (e.g., *c’est* [‘it’s’]) and blank spaces as multiple words.

The 'Custom list' feature allows users to create text profiles in any of the three languages using any word list they wish. If the user's custom list is organised into word families, the 'Word Family Statistics' table will work as normal. If the list is tagged by part of speech, the option to (de)select grammar features is also available. This functionality has proven useful to Awarding Organisations in the development of their draft GCSE word lists (e.g., AQA, 2023; Edexcel, 2023; Eduqas, 2022). As no data entered into the 'Custom list' field or elsewhere in the interface is sent to a server for processing (i.e., all processing happens locally), sensitive data remains completely secure, which is critical for commercially sensitive list preparations.

3.3. Embedding bespoke word lists: An example from the context of secondary school education in England

We have seen how frequency-based word lists organised by (f)lemma can be customised using *MultilingProfiler* to better reflect the word knowledge of a target group of beginner/low-intermediate learners. However, frequency-based profiling may still be too broad to be useful in materials creation for the very early stages of learning, when learners have very limited vocabularies that may be smaller than a few thousand words and perhaps only partial knowledge of grammatical paradigms. We give an example of this from the specific context in England. The new GCSE content (Department for Education, 2022) states that students should be able to recognise and produce at least 1,200 lexical items in the L2 at Foundation tier, and 1,700 at Higher tier,[3] 85% of which should be from the 2,000 most frequent words in the language. They also recommend including up to 30 multiword units and 20 culture-specific terms. Thus, any word list designed to align with the requirements of the GCSE is likely to contain (i) fewer than 2,000 headwords, (ii) a number of low frequency words relevant to the needs and interests of the target group, and (iii) some proper nouns and multiword units. So, a lexical profile created using frequency-based lists will not give very accurate representation of the words GCSE students can recognise, and a more bespoke list type is needed. We have embedded two such list types in *MultingProfiler*: curriculum-aligned lists which allow users to create bespoke text profiles using word lists aligned with the target knowledge expected at the end of the course (discussed in section 3.3.1) and cumulative lists that grow with learners' knowledge as they progress through the course (presented in section 3.3.2).

3.3.1. Curriculum-aligned lists

Currently, *MultilingProfiler* provides the option to profile texts using 'Eduqas/LDP GCSE' lists which align with the parameters set out in the new GCSE content. The lists were developed jointly by the Welsh Joint Examining Council's Awarding Organisation 'Eduqas' in collaboration with the 'Language-driven Pedagogy' (LDP) project (formerly the National Centre for Excellence for Language Pedagogy [NCELP][4]) (see Finlayson et al., under review, for the list creation methodology). For each language, users have the option to toggle between the tiers of entry for the GCSE (Foundation and Higher) and the modalities (reading and listening). The listening lists count bespoke (f)lemmas, defined as per the grammatical specifications for each tier in the curriculum content. For example, the (f) lemmas in the Spanish Foundation-tier lists contain only singular forms of verbs in the imperfect and inflectional future tenses, whereas the (f)lemmas in the Higher-tier lists also include plural forms. The reading lists add to these bespoke families the derivational patterns specified for each tier (Department for Education, 2022) on the grounds that there is more time to process written input than oral input, and that knowledge of sound-spelling relations is less likely to hinder recognition (e.g., the *-able* suffix in Spanish and French is identical to English in its orthographical form, but not its phonological form).

At the time of writing, collaboration is underway with two other major Awarding Organisations in England and Wales (AQA and Edexel) to embed curriculum-aligned word lists that support preparation for the new GCSE examinations (for teaching from 2024).

3.3.2. Cumulative lists

Text profiles created using curriculum-aligned lists represent the compatibility of a text with learners' expected knowledge at the *end* of a program of study. However, such lists are not sensitive to learners' development as they progress through a curriculum. The 'NCELP KS3' and 'NCELP KS4' lists embedded in *MultilingProfiler* showcase how cumulative profiling aligned with stages in a program of study can be achieved using word lists that grow in breadth (as learners' vocabulary size increases) and depth (as lexical and inflectional morphological knowledge develops). The programs of study with which these lists are aligned are the NCELP schemes of work developed to support (pre)-GCSE teaching of French, German, and Spanish in secondary schools in England (NCELP, 2021; 2022b). These schemes of work define the course structure and content (vocabulary, grammar and phonics) for teaching across Key Stages 3 (KS3) and 4 (KS4)[5] (Years 7–11; ages 11–16). Teachers and materials developers following the NCELP schemes of work can select any year, term, and week in the program and create text profiles that align with the vocabulary and grammar knowledge expected at that stage (see Fig. 4). That is, this functionality generates lexical profiles that reflect the language that has been included in intentional learning activities up to a specific point in a curriculum. From Year 10, users can also toggle between Foundation and Higher-tier versions of the lists.

Bespoke word lists are provided for every week in the NCELP scheme of work, with the content of each week's list building on that representing the previous week. Table 1 shows the number of headwords students have encountered, and, therefore, the length of the

[3] GCSE students are entered to sit examinations at either Foundation or Higher tier. At Higher tier, passing students are awarded a mark in the grade bracket 4–9, where 9 is the highest possible grade. At Foundation tier, the bracket is 1–5, with a grade of 4 required to pass.

[4] Department for Education funding for the National Centre for Excellence for Language Pedagogy ceased on 2 March 2023. Since then, the project continues to operate under the name 'Language-driven Pedagogy'.

[5] Key Stages 3 (Years 7–9, ages 11–14) and 4 (Years 10–11, ages 14–16) are stages of secondary education in England and Wales. Study of an MFL is obligatory at KS3. About half of students choose to study an MFL at KS4 and take the GCSE.

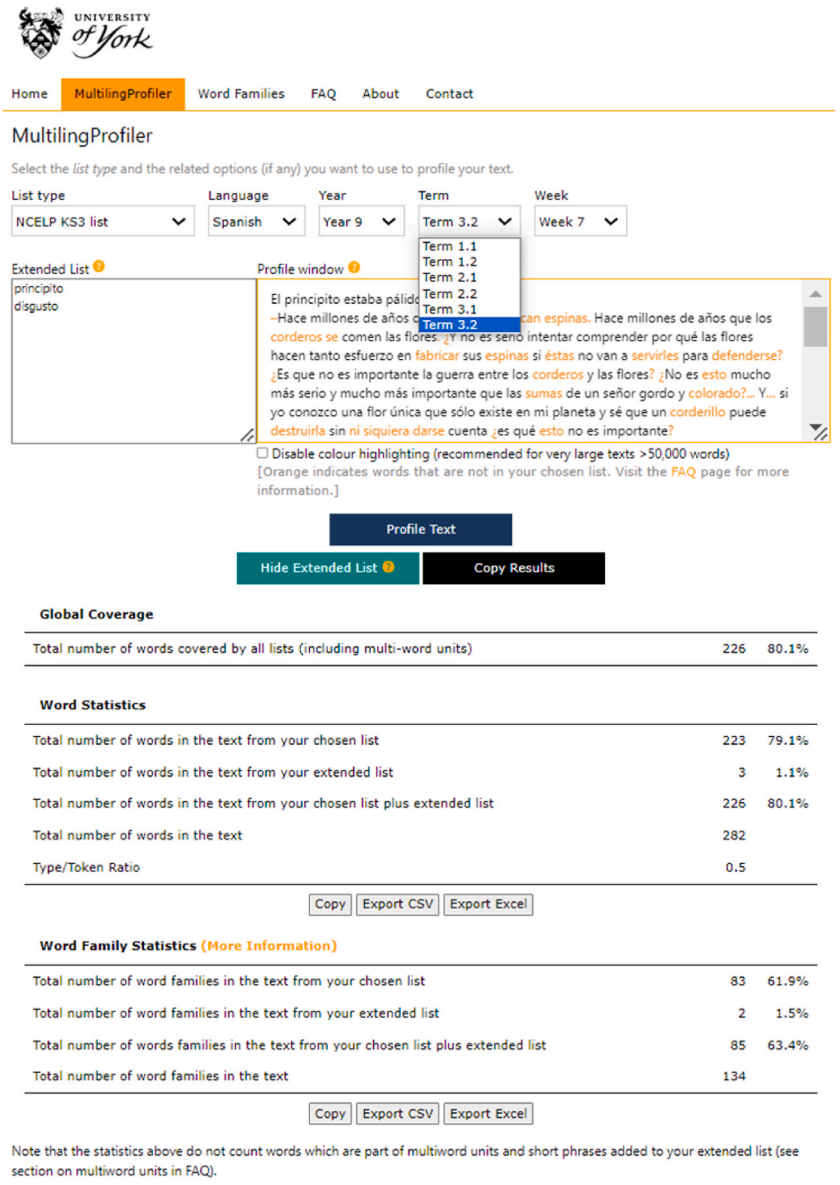

Fig. 4. Profile for *El Principito* at Year 9, Term 3.2, Week 7 in the NCELP Spanish scheme of work.

Table 1
Breadth of word knowledge across NCELP German KS3 (Year: Term).

	7 : 1	7 : 2	7 : 3	8 : 1	8 : 2	8 : 3	9 : 1	9 : 2	9 : 3
Headwords	145	256	354	559	643	730	822	868	950

word lists, at the end of various stages in KS3 German.

Figs. 5–7 illustrate how the number of word family members on the cumulative lists increases as more inflectional morphological patterns are introduced across KS3. The word forms are tagged with the following information about the point in the scheme of work at which they are introduced: the school year (7, 8, or 9); the term (1, winter; 2, spring; or 3, summer); the half term (1st or 2nd); the week in the half term (1st -7th). So, word forms with tag 7.1.2.4 are introduced in the fourth week of the second half of the first term in Year 7. Fig. 5 shows the new vocabulary headwords introduced in week 7.1.2.4 in the NCELP (2021) German scheme of work, which pertain to the topic of asking and answering questions about classroom scenarios: *lesen* ('to read'), *sprechen* ('to speak'), *wiederholen* ('to repeat'), *zeigen* ('to show'), *zuhören* ('to listen'), *Antwort* ('answer'), and *freiwillig* ('voluntary'). At the point of introduction, the size of the word families for verbs is restricted to four forms of the regular verbs *wiederholen* and *zeigen* (infinitive and first, second, and third

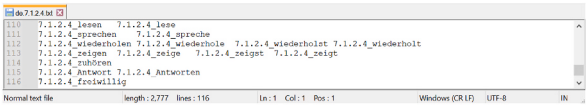

Fig. 5. Word family sizes at Year 7, Term 1.2, Week 4 in the NCELP German Scheme of Work.

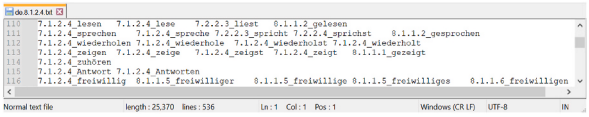

Fig. 6. Word family sizes at Year 8, Term 1.2, Week 4 in the NCELP German scheme of work.

Fig. 7. Word family sizes at Year 9, Term 1.2, Week 4 in the NCELP German scheme of work.

person singular), two forms of the stem-changing verbs *lesen* and *sprechen* (infinitive and third person singular), and one form of the separable verb *zuhören* (infinitive). By the same point in Year 8, students have further developed their knowledge of inflectional morphemes, and the word families now include the second and third person singular forms of stem-changing verbs (*liest, spricht, sprichst*) as well as the past participles (*gelesen, gesprochen, gezeigt*). By Year 9, the second person plural form of stem-changing verbs (*lest, sprecht*) has also been taught, and so the word families for this type of verb have grown in size again. The word family for *freiwillig* has also grown, following the teaching of adjective declension patterns.

From the first week of KS4 (10.1.1.1), NCELP provide different schemes of work for Foundation tier and Higher tier. These align with the vocabulary and grammar specified for each in the GCSE subject content (Department for Education, 2022). Thus, the size of word families can also differ, at any one point in time, between the lists embedded for each tier. Figs. 8 and 9 show the difference in size between some German word families first introduced towards the end of Year 7 by the time Foundation-tier and Higher-tier students respectively reach Week 6 in Term 1 of Year 10. In these examples, the simple past forms of the verbs *dauern* ('to last'), *erreichen* ('to reach'), *schaffen* ('to manage'), and *suchen* ('to look for')' and the dative plural form of the noun *Land* ('country') are included in the bespoke word family for Higher tier, but not in the bespoke word family for Foundation tier (there is no difference between the sizes of the families for other headwords in the set, which are [proper] nouns and adverbs). Such differences have the potential to significantly influence the lexical coverage of texts provided by the Higher list compared to its Foundation equivalent, as we will see in Illustrative Case Study 2.

Word lists for cumulative profiling are somewhat time consuming to develop, as word families must be manually created for each week and subsequently extended each time a relevant grammar paradigm is introduced. However, the benefits are a more precise and learner-centred approach to profiling that supports bespoke materials creation for a unique course of study, where such a curriculum design and pedagogical approach is appropriate. Of course, a tightly structured, meticulous approach to language growth and curriculum design may not be relevant in all educational contexts, such as those where more emphasis is given to strategies like inferencing or more generous assumptions are made about learners' capacity to recognise full lemmas in early stages of learning, for example purely analytic syllabi such as task-based or immersion approaches (see e.g., Long & Crookes, 1992). However, with synthetic syllabi (i.e., language-driven curricula), where linguistic content is planned and sequenced to accumulate over time, it can be useful to

Fig. 8. Word family sizes at Year 10, Term 1.1, Week 6 in the NCELP German Foundation-tier scheme of work.

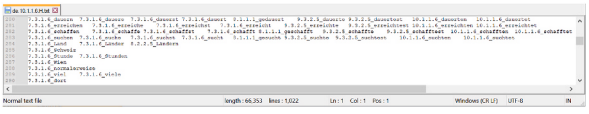

Fig. 9. Word family sizes at Year 10, Term 1.1, Week 6 in the NCELP German Higher-tier scheme of work.

reflect this growth in the pedagogical materials themselves. With this in mind, *MultilingProfiler* has been designed to accommodate other similar approaches in future, such as cumulative profiling on a 'year-by-year,' 'unit by unit' or 'task-by-task' basis.

4. Applications of *MultilingProfiler*: Illustrative case studies

We now evaluate the usefulness of our proposed contributions to more nuanced lexical profiling in French, German, and Spanish in two illustrative cases studies that showcase how *MultilingProfiler* may be used in the practical development of research-informed pedagogical materials for these languages. In section 4.1, we explore whether use of the 'Remove Inflected Forms' and 'Add Derived Forms' functions developed in response to our first aim may create more representative profiles of texts intended for use with low-intermediate learners of French. In section 4.2, we trial the use of cumulative profiling, one of our approaches to operationalising our second aim, to situate texts at an appropriate stage and level in a German course.

The general principle behind the type of lexical profiling research used in each study is to determine how many off-list words appear in a text, and consider the learning and/or assessment opportunities such words provide (see Strand C in Nurmukhamedov & Webb, 2019, for an overview of other lexical profiling studies of this type). Alignment between what is taught and assessed is known to be important for test validity and reliability (e.g., Nation, 2022; Schmitt, 2000), especially in the case of vocabulary achievement tests that assess knowledge of a specific set of words (Nation, 2022) or vocabulary skills (Read, 2000) taught in a course (as opposed to proficiency tests of comprehension and production).

English glosses of all example texts used in this section can be found in Appendix D.

4.1. Illustrative case study 1: Developing graded materials with bespoke word frequency lists

Lexical profiling with frequency-informed word lists is used widely in the development of graded materials for extensive listening and reading (e.g., Dang, 2023; Hill, 2001; Nation & Deweerdt, 2001; Nation & Waring, 2019). These levelled texts provide accessible target language input for learners who are assumed to have mastered the words in a certain frequency band, with the aim of enhancing their lexis and improve their reading speed (Hill, 2001). This approach is useful when there are general expectations about vocabulary size and depth, for example, in the development of CEFR-classified materials, examinations, and placement tests.

As we have seen, this process might not always be as simple as taking the raw coverage value provided by a frequency band as an indication of the comprehensibility of a text written in an inflected or declined language. For example, it might be reasonable to assume that a candidate with English as an L1 preparing to take the *Diplôme d'études en langue française* (DELF) examination at CEFR B1 level can recognise the 2,000 most frequent French words in writing, and that any text with at least 95% lexical coverage by the 2,000-band is well suited as a reading comprehension or translation task at this level (Laufer & Ravenhorst–Kalovski, 2010). At first glance, the extract in Fig. 10 from Perrault's *Cendrillon* (Cinderella) would seem to fit this criterion. With proper nouns (*Cendrillon*) and cognates with English (*princesse, duchesse*) added to the extended list, the 2,000 band gives coverage of 92.9% of the words in this text. Glossing the recurrent key word *pantoufle* ('slipper') would bring the figure to 96.4%, i.e., within the recommended margins for adequate comprehension.

However, while mastery of the 2,000 most frequent lemma headwords might be a reasonable expectation of a B1-level student, the grammatical knowledge necessary to recognise a number of complex forms which are part of French lemmas cannot be assumed at this level. When the past historic tense and the present and imperfect subjunctive moods are deselected in the 'Remove Inflected Forms' dropdown, the text coverage decreases to 86.4%, i.e., well below the recommended comprehension margin (see Fig. 11). The majority of words that appear orange (off list) following the removal of these forms are past historic or imperfect subjunctive forms of highly frequent verbs (e.g., *fit* and *firent* from *faire* ['to make, do']; *purent* from *pouvoir* ['to be able to']; *mit* and *mirent* from *mettre* ('to put'); *fit* and *fût* from *être* ['to be']). As these highly irregular forms (i) bear little or no resemblance to the headword stem and are, thus, unlikely to be intuitively recognised as part of its family, and (ii) are crucial to the understanding of the sentences in which they appear, their

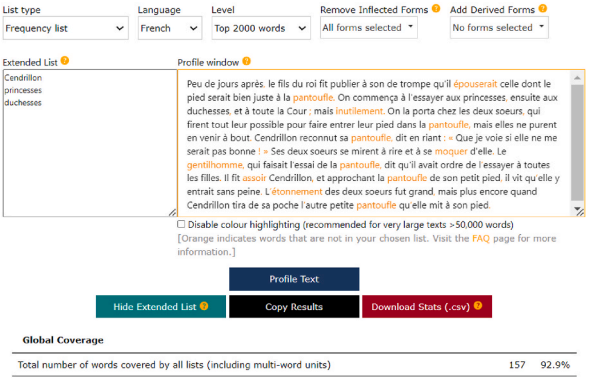

Fig. 10. Profile for *Cendrillon* sample against 2,000 band before removing inflected forms.

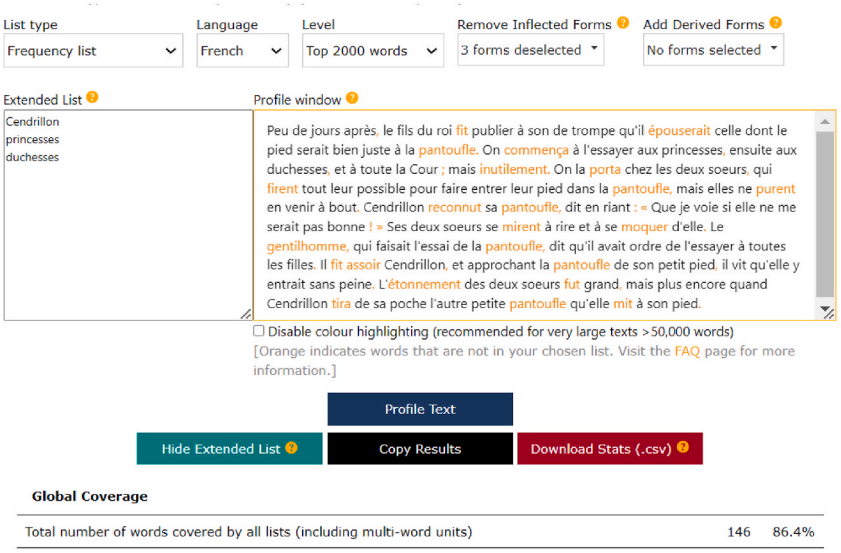

Fig. 11. Profile for *Cendrillon* sample against 2,000 band after removing inflected forms.

presence is likely to pose an obstacle to comprehension at this level. This may be desirable, if the planned use of the text is to teach irregular verbs in the past historic and imperfect subjunctive or to practise strategies such as inferencing. However, if material for a comprehension or translation activity is sought, this text may not be suitable in its current form. If desired, materials developers can use the editing function of *MultilingProfiler* to adapt the text, for example by glossing the irregular verbs and adding them to the extended list, replacing the irregular forms with (more) regular items, or rewriting the text in the perfect tense (a more frequently used, non-literary past tense in French which expresses the same meaning as the past historic).

Conversely, B1-level students may be reasonably expected to comprehend frequent, productive, predictable, and regularly derived forms with *in-/im-* and *-ment* affixes. Fig. 12 shows the small overall increase in coverage when these affixes are selected in the 'Add Derived Forms' dropdown.

This illustrative case study highlights the strengths of *MultilingProfiler* for frequency-based profiling of French, German, and Spanish texts intended for use with beginner and low-intermediate learners. While we have seen that there may be less need for tools to be sensitive to grammatical inflection in texts written for learners of English, such a feature seems critical to success when processing texts in French. Had this text been profiled using a tool that did not allow users to deselect grammar features, the comprehensibility of the text for a low-intermediate learner of French may have been overestimated.

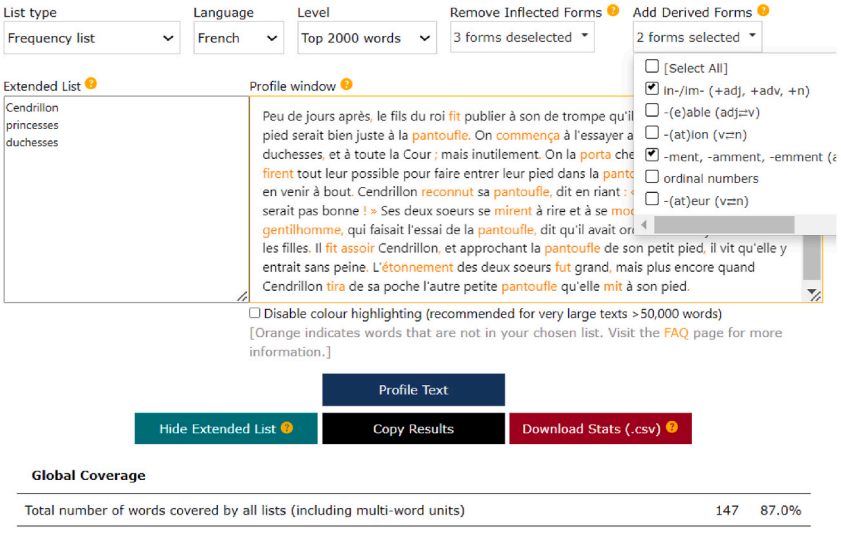

Fig. 12. Profile for *Cendrillon* sample against 2,000 band after removing inflected forms and adding derived forms.

4.2. *Illustrative case study 2: Using cumulative word lists to situate materials in a program of study*

In section 3.3.2, we saw how word lists based on curriculum stage can grow substantially in breadth and depth as learner knowledge accumulates over time. Here, we use the cumulative word lists embedded in *MultilingProfiler* to assess the extent to which this growth affects the lexical coverage of text provided by lists aligning with different points in the NCELP German schemes of work (NCELP, 2021; 2022b), and situate two example texts at appropriate stages in the program accordingly.

The word lists used to create the three profiles in Figs. 13–15 represent the totality of the vocabulary and grammar content covered by the schemes of work for Years 7, 8, and 9 respectively. The sample text is part of a GCSE German Foundation-tier reading paper (AQA, 2020: 15). The profiling output (67.7% coverage) shows that the text is unlikely to be appropriate for use at the end of Year 7, at which point learners have not yet been introduced to the vocabulary or grammar knowledge necessary for reliable comprehension (the orange words mainly comprise function words, accusative and dative forms of articles and pronouns, prepositions, reflexive pronouns, and simple past and subjunctive forms of verbs). At the end of Year 8 (89.2% coverage), the text could feasibly serve as a dictionary or inferencing exercise, as the remaining orange (off list) words comprise mainly content words. Alternatively, it could provide a useful introduction to the subjunctive verb form *möchte* ['I/s/he would like']). By the end of Year 9, the profiling output (94.6% coverage) suggests that the majority of the vocabulary and grammar necessary to adequately comprehend the text has been introduced, and students should be able to cope with the lexical demands of the text. The differences in lexical coverage provided by lists representing the vocabulary knowledge introduced in Years 7, 8, and 9 have a significant effect on the overall usefulness and potential purposes of text at different stages in the KS3 curriculum. Without a lexical profiling tool, an awareness of the suitability of the text and the words that might need additional attention at different stages will not be accessible to anyone other than a very experienced and attentive user of the NCELP schemes of work.

MultilingProfiler can also be used to assess the suitability of texts for learner groups who have had the same amount of exposure to instruction, but progress at different rates (i.e., they are working with different language systems at any given point in time). Figs. 16 and 17 show profiles of an example text created using Higher-tier and Foundation-tier versions of the same German word list (NCELP, 2022b). The text is taken from a portfolio of preparatory materials for a German examination at CEFR A2-level (Goethe-Institut, 2016: 6–7). The Higher-tier version of the word list gives substantially greater lexical coverage of this text than its Foundation-tier counterpart, largely due to the inclusion of simple past verb forms in the prescribed language content for the Higher tier, but not the Foundation tier (Department for Education, 2022). Glossing key words *Sendung* ('programme') and *Koch* ('cook') increases coverage by the Higher-tier list to 95%, suggesting that the text in this activity should be comprehensible for learners if translations of these two terms are provided. The coverage by the Foundation-tier list, however, is just 79%, suggesting that learners working at this level are likely to struggle with the lexical load of this text in its current form. An adapted version could be created for Foundation-tier students by replacing the verbs in simple past with perfect tense forms (a more frequently used past tense in German which expresses the same meaning as the simple past). Again, although there is a major difference in the degree of coverage provided by the Foundation-tier and Higher-tier lists, the extent and nature of the differences are unlikely to be recognised intuitively by anyone without a very in-depth knowledge of the GCSE subject content.

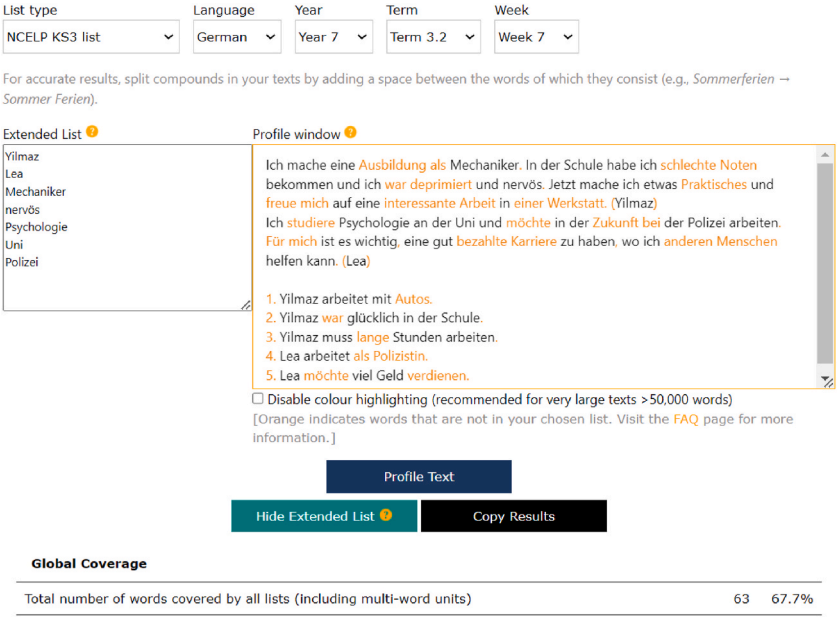

Fig. 13. Profile for GCSE German Foundation-tier reading paper sample at Year 7, Term 3.2, Week 7 in the NCELP scheme of work.

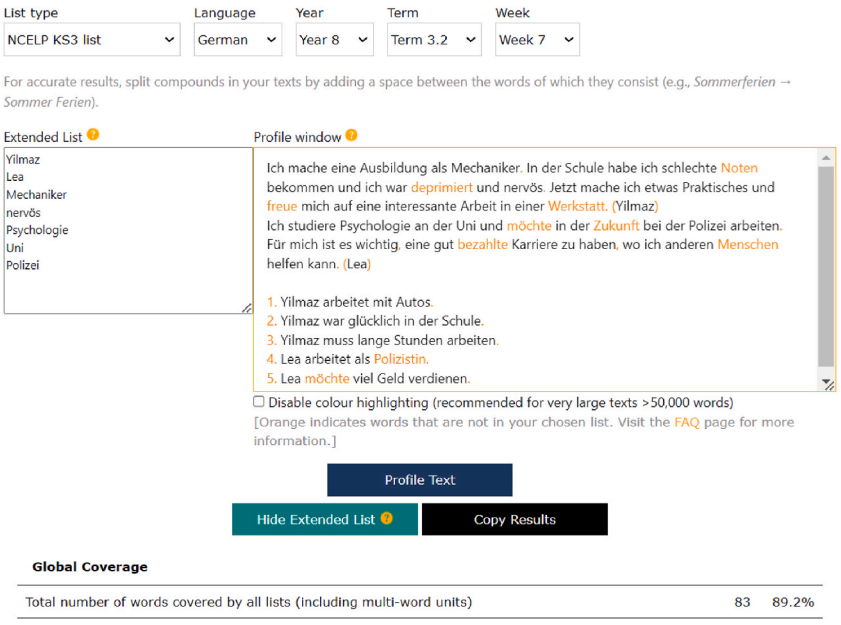

Fig. 14. Profile for GCSE German Foundation-tier reading paper sample at Year 8, Term 3.2, Week 7 in the NCELP scheme of work.

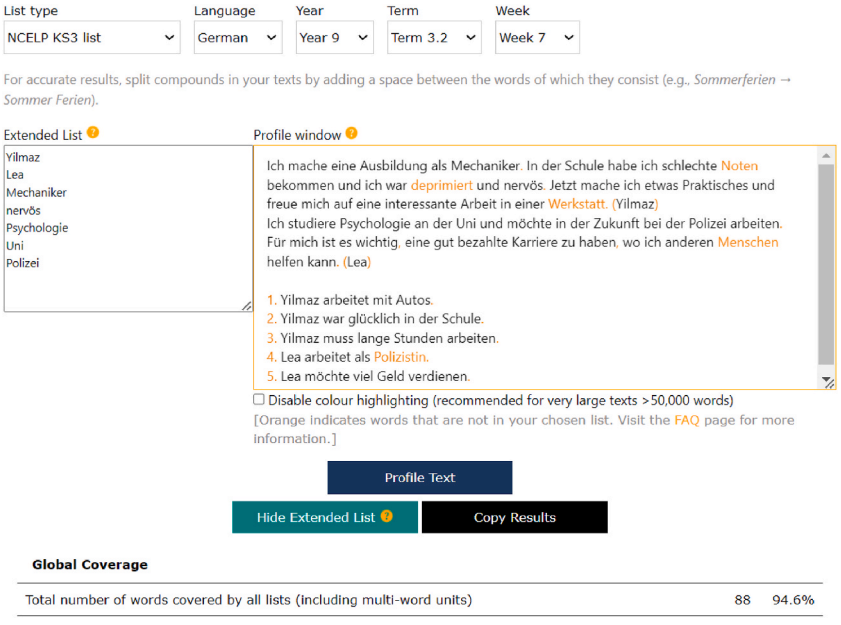

Fig. 15. Profile for GCSE German Foundation-tier reading paper sample at Year 9, Term 3.2, Week 7 in the NCELP scheme of work.

In sum, we have seen how two key features of *MultilingProfiler* designed to enable more bespoke lexical profiling of texts in French, German, and Spanish—adapting (f)lemmatised word frequency lists by removing inflected forms and adding derived forms, and profiling with embedded, cumulative word lists that grow with learner knowledge—can provide insights into the suitability of texts for different learner groups that may be overlooked if use of fully (f)lemmatised lists or intuitions about learner knowledge at a particular stage of development are relied upon.

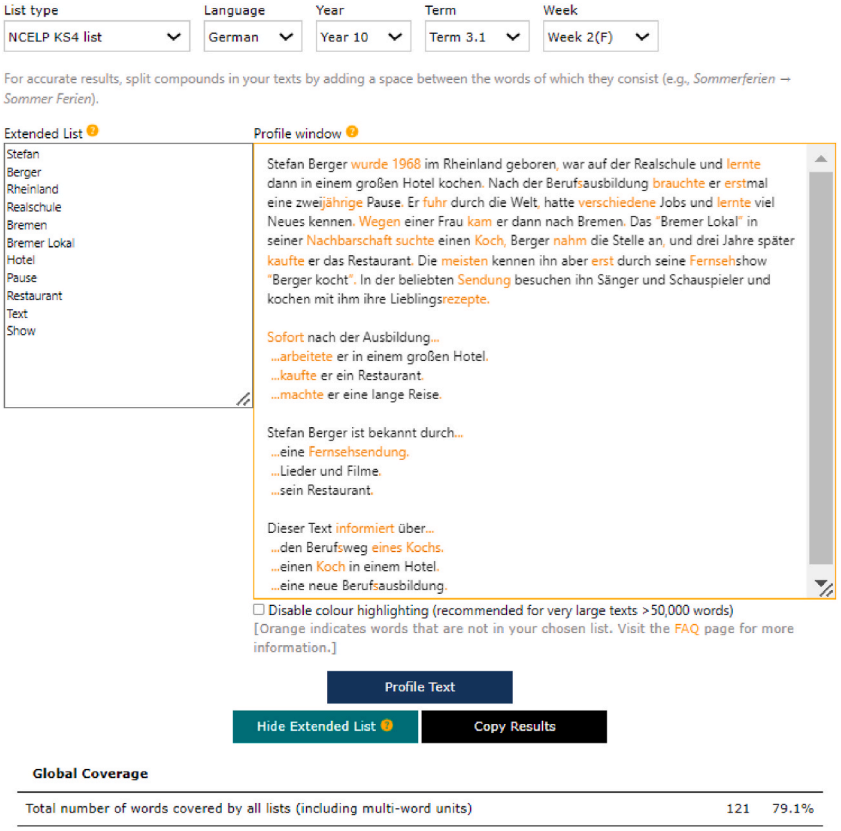

Fig. 16. Profile of A2 German reading materials at Year 10, Term 3.1, Week 2 in the Foundation-tier version of the NCELP KS4 scheme of work.

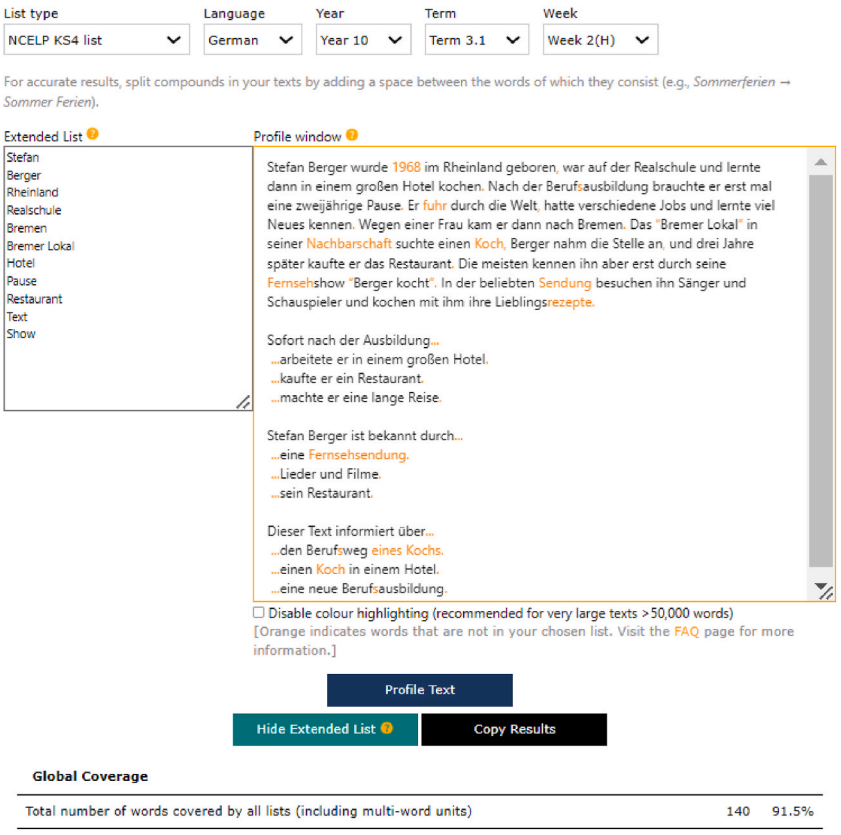

Fig. 17. Profile of A2 German reading materials at Year 10, Term 3.1, Week 2 in the Higher-tier version of the NCELP KS4 scheme of work.

5. Limitations of *MultilingProfiler* and future directions

Currently, the most significant limitation to the accuracy of text profiles created using *MultilingProfiler* is the lack of a part-of-speech tagging or semantic tagging functionality, which means that polysemous and homonymous word forms in the profiling output are not disambiguated for part of speech or sense. For example, no distinction is made between the Spanish prepositional *para* ('for, in order to') and verbal *para* ('s/he stops'). When calculating word family statistics, *MultilingProfiler* assigns polysemous and homonymous forms to the word family headword which appears first on a given word list (e.g., for a frequency-ordered list, it assumes that *para* is part of the prepositional family with frequency ranking 16, rather than the verbal family with frequency 706). We are currently exploring options for part-of-speech tagging that do not dramatically reduce the speed of local processing. Addressing the issue of semantic polysemy is a greater challenge. As semantic tagging systems for these languages are in development and generally operate with relatively low accuracy, no plans are currently in place to introduce tagging by semantic category. Thus, polysemous and homonymous words with the same part of speech (e.g., *la livre* ['pound'] and *le livre* ['book']) will continue to be counted together, and human checking is needed to resolve these ambiguities.

A set of limitations specific to lexical profiling in German lies in the complexities introduced by separable verbs and compounding in this language. *MultilingProfiler* recognises infinitive and non-separated forms of German separable verbs (e.g., *aufstehen, aufstehe* (['to stand up, I stand up']) if they are on the word list, but users must add the base verbs and prefixes in their separated forms (e.g., *stehe* ['I stand'], *auf* ['up']) to the extended list to include them in the profile (in cases where those forms do not appear on the list in their own right). The development of compound-splitting software to make German texts more compatible with lexical profiling approaches is ongoing (e.g., Tuggener, 2016). For now, when German is selected as the target language, a pop up appears reminding users to split any compounds manually for more accurate results.

Work to address these limitations will certainly increase the potential for *MultilingProfiler* to be used as a research tool. Nurmukhamedov and Webb (2019) compiled a timeline of lexical profiling research in language education, which they organised into four major types of research: (i) experimental studies with participants investigating the amount of lexical coverage needed to reach adequate comprehension of different types of discourse; (ii) research with corpora and word lists to investigate the number and type of words needed to reach the coverage figures associated with adequate comprehension; (iii) textual analysis to investigate opportunities for (incidental) vocabulary learning and practice provided by different text types and scenarios; (iv) methodological research about tools and approaches (of which this article is an example). Indeed, the current features of *MultilingProfiler* support small-scale research of any of these types, and several studies using the tool are already underway. Examples of type (ii) research include studies that look at

the lexical content of French, German, and Spanish GCSE examination papers (Dudley & Marsden, 2023) and analyse the relative coverage of these examination papers by Awarding Organisations' existing, topic-based word lists compared with the new frequency-informed lists (Finlayson et al., under review; Marsden et al., 2023). An example of type (iii) research is (Mitchell & Myles, 2023) study of the frequency of vocabulary used in teacher input in a French primary classroom. We are aware of one use of *MultilingProfiler* in type (i) research (albeit to look at production rather than comprehension) on the development of the lexical richness of texts written by beginner learners of French (Vold, 2022), which is encouraging as we have seen that studies of this type with learners of French, German, and Spanish are few. We believe that *MultilingProfiler* can play a role in facilitating this much needed research in the future by enabling researchers to easily create experimental texts with different vocabulary levels. Such a technique has a place in the creation of research-informed intervention studies and tests for non-English languages, including inferencing research (e.g., Laufer, 2020) and incidental learning studies (e.g., Godfroid et al., 2013).

Longer-term plans to develop *MultilingProfiler* for both pedagogical and research purposes include expanding the 'Add Derived Forms' feature (by affixes at levels 4-7 and making the feature available for headwords in frequency bands 3,000–5,000), adding the capacity to profile in other languages, and developing a function to profile multiple texts simultaneously (corpora). The organisation of training events for teachers on use of the tool and interpretation of results is also envisaged. A particularly valuable aspect of user-friendly, freely accessible profiling tools is their capacity to empower teachers to create bespoke, research-informed materials tailored to the needs and interests of their classes. Indeed, we are aware of practising teachers in the NCELP network of schools who are using this cumulative profiling function to write and adapt texts to suit the needs of their individual classes, having undertaken training in the method as part of CPD provided by NCELP.

6. Conclusions

In this article, we established a need for tools and word lists that can manage the complexities of textual analysis in grammatically complex languages, particularly where such texts are intended for use with learners working at beginner and low-intermediate proficiency levels. We introduced the concept of 'bespoke word families', and developed *MultilingProfiler* to operationalise this concept with two key aims in mind: (i) to develop a vocabulary profiling tool that supports the flexible profiling of texts in French, German, and Spanish, and (ii) to develop an approach to bespoke word list creation that supports the profiling of texts in French, German, and Spanish for particular learner groups.

To address the first aim, we developed toggles that allow users to adapt default (f)lemmatised lists by removing complex inflected forms and adding straightforward derived forms. We tested the usefulness of this feature in an assessment of the suitability of a literary text for a CEFR B1-level learner, and found that default (f)lemmas may be an unreliable unit of counting for this purpose. We also described the development of word boundary markers that account for variation in orthographic rules across languages, and the option to add multiword unit families to word lists in recognition of the role chunks play in early-stage language education. In response to the second aim, we created curriculum-based word lists to support preparation for GCSE qualifications in French, German, and Spanish, and cumulative versions of these that grow as learners progress through the program and accumulate knowledge on a week-by-week basis. We used this feature to situate two GCSE-level texts at appropriate points and tiers in the program of study, and saw how *MultilingProfiler* is effective in ensuring alignment between what has been taught and what can be tested to a degree of detail that would be difficult to achieve intuitively.

More generally, we have described *MultilingProfiler*'s ability to calculate lexical coverage of texts at the level of the word, token, (with types calculable by multiplying tokens by the type/token ratio provided), (f)lemma, word family level 3, or bespoke word family, and to offer detailed statistical information about the use of words, word forms, and multiword units. Users can customise pre-existing lists using the 'extended list' function, edit texts directly in the user-friendly profile window, and immediately reprofile them. We have also presented examples of how *MultilingProfiler* has been used for research purposes, and pointed to future developments that should expand the scope of the tool in research and education.

Author statement

Natalie Finlayson: Conceptualisation**,** Methodology, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualisation, Project administration. **Emma Marsden:** Conceptualisation**,** Methodology, Validation, Formal Analysis, Funding acquisition, Investigation, Resources, Writing - Review & Editing, Project administration, Supervision. **Laurence Anthony:** Conceptualisation, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing - Review & Editing, Visualisation, Supervision.

Funding

This work was supported by The Department for Education (MFL Hubs, 30365), Research England (York, 2019), and ESRC (Impact Acceleration Account, York, 2018).

Acknowledgements

We would like to thank Inge Alferink, Nick Avery, Louise Bibbey, Giulia Bovolenta, Louise Caruso, Emily Cutts, Elin Glaves, Rachel Hawkes, Amanda Izquierdo, Heike Krüsemann, Catherine Morris, Ciarán Morris, Charlotte Moss, Stephen Owen, Jack Peacock,

Catherine Salkeld, Lauren Smith, and Peter Watson for their contributions to the development of *MultilingProfiler*, and the reviewers for their helpful comments that improved the quality of the paper.

Appendices.

Appendix A: Examples of inflectional richness in English, French, German, and Spanish

Table A.1
Size of determiner, noun, and adjective lemmas in English and German

Part of Speech	English	German
Determiner	a: a, an	*ein: ein, eine, einem, einen, einer, eines*
Noun	book: book, books	*Buch: Buch, Buches, Buchs, Bücher, Büchern*
Adjective	big: big, bigger, biggest	*groß: groß, große, großem, großen, großer, großes, größer, größere, größerem, größeren, größerer, größeres, größte, größtem, größten, größter, größtes*

Table A.2
Size of verb lemmas in English, French, and Spanish

Part of speech	English	French
irregular verb	be: be, am, are, is, being, been, were	*être: être, es, est, étaient, étais, était, étant, été, êtes, étiez, étions, fûmes, furent, fus, fusse, fussent, fusses, fussiez, fussions, fûtes, fut, fût, sera, seraient, serai, serais, serait, seras, serez, seriez, serions, serons, seront, soient, sois, soit, sommes, sont, soyez, soyons, suis*
		Spanish
regular verb	touch: touch, touches, touching, touched	*tocar: toca, tocaba, tocabais, tocaban, tocabas, tocad, tocada, tocadas, tocado, tocados, tocamos, tocan, tocando, tocara, tocarais, tocaran, tocaras, tocare, tocareis, tocaremos, tocaren, tocares, tocaron, tocará, tocarán, tocarás, tocaré, tocaréis, tocaría, tocaríais, tocaríamos, tocarían, tocarías, tocas, tocase, tocaseis, tocasen, tocases, tocaste, tocasteis, toco, tocábamos, tocáis, tocáramos, tocáremos, tocásemos, tocó, toque, toquemos, toquen, toques, toqué, toquéis*

Appendix B: List of lemmatised and part-of-speech-tagged languages in Sketch Engine

Afrikaans, Arabic, Bulgarian, Catalan, Crimean Tatar, Croatian, Czech, Danish, Dutch, English, Estonian, Filipino, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Icelandic, Irish, Italian, Japanese, Korean, Latvian, Norwegian (Bokmål and Nynorsk), Polish, Portuguese, Romanian, Russian, Serbian (including Latin), Slovak, Slovenian, Spanish, Swahili, Swedish, Tagalog, Tibetan, and Ukrainian.

Appendix C: (De)selectable inflected forms and derived forms in MultilingProfiler

Table C.1
(De)selectable inflected forms in each language

Language	Inflected forms
French	imperfect, past historic, inflectional future, past participle, conditional, present subjunctive, imperfect subjunctive, present participle, imperative
German	simple past, past participle, present participle, genitive noun, imperative, subjunctive
Spanish	preterite, imperfect, inflectional future, conditional, present subjunctive, imperfect subjunctive, future subjunctive, present participle, past participle, imperative, verbs with cliticised pronouns as suffixes

Notes. The present tense indicative cannot be deselected, and so is not listed.

Table C.2
(De)selectable derived forms in each language

Language	Affixes
French	in-/im- (+adj, +adv, +n), -(e)able (adj ⇄ v), -(at)ion (v ⇄ n), -ment, -amment, -emment (adj ⇄ adv), ordinal numbers, -(at)eur (v ⇄ n)
German	un- (adj ⇄ adj), Haupt- (n ⇄ n), Lieblings- (n ⇄ n), -ung (v ⇄ n), -er (v ⇄ n), ordinal numbers, -heit (adj/adv ⇄ n), -keit (adj/adv ⇄ n), -los (n ⇄ adj), -chen (n ⇄ n), -lein (n ⇄ n)
Spanish	-(a)mente (adj ⇄ adv), -idad (adj ⇄ n), -ísimo (adj ⇄ adj), -able (v ⇄ adj), -ito (n ⇄ n)

Notes. ⇄ indicates the pattern directionality; i.e., both derived forms of a base word and base forms of a derived word will be included in the word definition when this affix is selected.

Appendix D: *English glosses of sample texts*

1. *Gloss of texts in Figs. 10 and 11*

A few days later, the king's son announced to the sound of trumpets that he would marry the girl whose foot fit the slipper. First, it was tried on princesses, then on duchesses and the whole court; but in vain. It was taken to the two sisters, who did everything possible to get their feet into the slipper, but they couldn't. Cinderella recognised her slipper and said with a smile: "Let me see if it would be good for me!" Her two sisters started laughing and making fun of her. The gentleman doing the slipper test said he had orders to try it on all the girls. He got Cinderella to sit down, bringing the slipper towards her little foot. He saw that it went in easily. The two sisters were greatly astonished, and even more so when Cinderella took the other little slipper out of her pocket and put on her foot.

2. *Gloss of text in Figs. 13 and 14*

I'm doing an apprenticeship as a mechanic. I got bad grades at school and was dejected and nervous. Now I'm doing something practical and looking forward to interesting work in a workshop. (Yilmaz)

I study psychology at uni and would like to work in the police force in the future. It's important for me to have a well-paid career in which I can help other people. (Lea)

1. Yilmaz works with cars.
2. Yilmaz was happy at school.
3. Yilmaz had to work long hours.
4. Lea works as a policewoman.
5. Lea has to earn a lot of money.

3. *Gloss of text in Figs. 16 and 17*

Stefan Berger was born in Rhineland in 1968, went to secondary school, and then learned to cook in a big hotel. After his training, he initially needed a two-year break. He travelled the world, had different jobs, and learned many new things. Then he came to Bremen because of a woman. The *Bremer Lokal* in his neighbourhood was looking for a chef. Berger accepted the position, and bought the restaurant three years later. Most people know him primarily through his TV show *Berger kocht*. In the popular programme, singers and actors visit him and cook their favourite recipes with him.

Immediately after the apprenticeship …

… he worked in a big hotel.
… he bought a restaurant.
… he went on a long trip.

Stefan Berger is famous for …

… a TV programme.
… songs and films.
… his restaurant.

This text gives us information about …

… the career path of a chef.
… a chef in a hotel.
… a new type of vocational training.

References

Anthony, L. (2022a). *AntWordProfiler* (Version 2.0.1) https://www.laurenceanthony.net/software.

Anthony, L. (2022b). *TagAnt (Version 2.0.5)*. https://www.laurenceanthony.net/software.

AQA. (2020). Question paper (foundation): Paper 3 reading. https://filestore.aqa.org.uk/sample-papers-and-mark-schemes/2020/november/AQA-86683F-QP-NOV20.PDF. (Accessed 12 September 2022).

AQA. (2023). *GCSE French*. Retrieved from https://filestore.aqa.org.uk/resources/french/specifications/AQA-8652-SP-2024.pdf Accessed August 27, 2023.

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253–279. https://doi.org/10.1093/ijl/6.4.253

Bax, S. (2012). *Text Inspector. Online text analysis tool*. http://textinspector.com.

Billuroğlu, A., & Neufeld, S. (2007). *BNL 2709: The essence of English* (4th ed.). Rüstem Kitabevi.

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics, 36*(1), 1–22. https://doi.org/10.1093/applin/amt018

Brezina, V., & Gablasova, D. (2017). *LancsLex: English vocabulary analysis tool*. http://corpora.lancs.ac.uk/vocab.

Brezina, V., Hawtin, A., & McEnery, T. (2021). The written British national corpus 2014 – design and comparability. *Text & Talk, 41*(5–6), 595–615. https://doi.org/10.1515/text-2020-0052

Brown, D., Stoeckel, T., Mclean, S., & Stewart, J. (2020). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics, 43*(3), 596–602. https://doi.org/10.1093/applin/amaa061
Browne, C. (2014). A new general service list: The better mousetrap we've been looking for? *Vocabulary Learning and Instruction, 3*(2), 1–10. https://doi.org/10.7820/vli.v03.2.browne
Cobb, T. (n.d.). VocabProfilers. https://www.lextutor.ca/vp.
Cobb, T., & Laufer, B. (2021). The nuclear word family list: A list of the most frequent family members, including base and affixed words. *Language Learning, 71*(3), 834–871. https://doi.org/10.1111/lang.12452
Coxhead, A. (2000). A new academic word list. *Tesol Quarterly, 34*(2), 213–238. https://doi.org/10.2307/3587951
Curcin, M., & Black, B. (2019). Investigating standards in GCSE French, German, and Spanish through the lens of the CEFR. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/844034/Investigating_standards_in_GCSE_French_German_and_Spanish_through_the_lens_of_the_CEFR.pdf. Accessed October 12, 2022.
Dang, T. N. Y. (2019). Corpus-based word lists in second language vocabulary research, learning, and teaching. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 288–303). Routledge.
Dang, T. N. Y. (2022). *Word Part Instructions in Pre-university EFL Contexts*. In *Presentation at the annual conference of the BAAL Vocabulary Special Interest Group*. Exeter, UK.
Dang, T. N. Y. (2023). Using VocabProfilers to select texts for extensive reading activities. In V. Viana (Ed.), *Teaching English with corpora: A resource book* (pp. 69–73). Routledge.
Davies, M., & Davies, K. (2019). *A frequency dictionary of Spanish: Core vocabulary for learners* (2nd ed.). Routledge.
Department for Education. (2022). GCSE modern foreign languages (MFL) subject content review. Retreived from https://www.gov.uk/government/consultations/gcse-modern-foreign-languages-mfl-subject-content-review. Accessed September 12, 2022.
Dijkstra, A., Miwa, K., Brummelhuis, B., Sappelli, M., & Baayen, H. (2010). How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language, 62*(3), 284–301. https://doi.org/10.1016/j.jml.2009.12.003
Dudley, A., & Marsden, E. (2023). The number and frequency of words 16-year-olds need for their French, German, and Spanish exams. *OSF Preprints*. https://doi.org/10.31219/osf.io/pzqkm.
Edexcel. (2023). GCSE (9-1) French specification - issue 1. Retrieved from https://qualifications.pearson.com/content/dam/pdf/GCSE/French/2024/specification-and-sample-assessments/gq000023-gcse-french-specification-2024-issue-1-1.pdf. Accessed August 27, 2023.
Eduqas. (2022). WJEC Eduqas GCSE (9-1) in French. Retrieved from https://www.eduqas.co.uk/umbraco/surface/blobstorage/download?nodeId=43422. Accessed July 7, 2023.
Estivalet, G. L., & Meunier, F. E. (2015). Decomposability and mental representation of French verbs. *Frontiers in Human Neuroscience, 9*(4), 1–10. https://doi.org/10.3389/fnhum.2015.00004
Finlayson, N., Marsden, E., & Anthony, L.. MultilingProfiler (Version 3). https://www.multilingprofiler.net/.
Finlayson, N., Marsden, E., & Hawkes, R. (Under review). A new vocabulary list for beginner-low-intermediate learners of French, German and Spanish aged 11-16.
Gardner, D. (2013). *Exploring vocabulary: Language in action*. Routledge.
Garner, J. (2022). VocabKitchen. http://vocabkitchen.com/profiler/.
Giordano, M. J. (2021). Lexical coverage in dialogue listening. *Language Teaching Research*. https://doi.org/10.1177/1362168821989869
Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye tracking. *Studies in Second Language Acquisition, 35*(3), 483–517. https://doi.org/10.1017/S0272263113000119
Goethe-Institut. (2016). Goethe-Zertifikat A2 Modellsatz Erwachsene. Retrieved from https://www.goethe.de/pro/relaunch/prf/materialien/A2/A2_Modellsatz_Erwachsene.pdf. Accessed September 12, 2022).
Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). RANGE and FREQUENCY programs. http://www.vuw.ac.nz/lals/staff/Paul_Nation.
Herman, E., & Leeser, M. J. (2022). The relationship between lexical coverage and type of reading comprehension in beginning L2 Spanish Learners. *The Modern Language Journal, 106*(1), 284–305. https://doi.org/10.1111/modl.12761
Hill, D. R. (2001). Survey review: Graded readers. *English Language Teaching Journal, 55*(3), 300–324. https://doi.org/10.1093/elt/55.3.300
Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430. https://doi.org/10.26686/wgtn.12560354
Iwaizumi, E., & Webb, S. (2022). To what extent do learner- and word-related variables affect production of derivatives? *Language Learning, 73*(1), 301–366. https://doi.org/10.1111/lang.12524
Kempley, S. T., & Morton, J. (1982). The effects of priming with regularity and irregularity related words in auditory word recognition. *British Journal of Psychology, 73*(4), 441–454. https://doi.org/10.1111/j.2044-8295.1982.tb01826.x
Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., … Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9
Laufer, B. (2020). Lexical coverages, inferencing unknown words and reading comprehension: How are they related? *Tesol Quarterly, 54*(4), 1076–1085. https://doi.org/10.1002/tesq.3004
Laufer, B. (2021). Lemmas, flemmas, word families and common sense. *Studies in Second Language Acquisition, 43*(5), 965–968. https://doi.org/10.1017/S0272263121000656
Laufer, B., & Ravenhorst–Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15–30.
Laufer, B., Webb, S., Kim, S. K., & Yohanan, B. (2021). How well do learners know derived words in a second language? The effect of proficiency, word frequency and type of affix. *ITL - International Journal of Applied Linguistics,, 172*(2), 229–258. https://doi.org/10.1075/itl.20020.lau
Long, M. H., & Crookes, G. (1992). Three approaches to task-based syllabus design. *Tesol Quarterly, 26*(1), 27–56. https://doi.org/10.2307/3587368
Lonsdale, D., & Le Bras, Y. (2009). *A frequency dictionary of French: Core vocabulary for learners*. Routledge.
Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22*(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov
Marsden, E., Dudley, A., & Hawkes, R. (2023). Use of word lists in a high-stakes, low-exposure context: Topic-driven or frequency-informed. *The Modern Language Journal*. https://doi.org/10.1111/modl.12866
Marslen-Wilson, W. D., Hare, M., & Older, L. (1995). Priming and blocking in the mental lexicon: The English past tense [Conference Presentation]. In *Presentation at the meeting of the Experimental Psychology Society*. London, UK.
McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics, 39*(6), 823–845. https://doi.org/10.1093/APPLIN/AMW050
Meunier, F., & Marslen–Wilson, W. (2004). Regularity and irregularity in French verbal inflection. *Language and Cognitive Processes, 19*(4), 561–580. https://doi.org/10.1080/01690960344000279
Mitchell, R., & Myles, F. (2023). Lexical input in the primary languages classroom: Frequency and appropriacy. In *Presentation at the BAAL Annual Conference 2023*. York, UK.
Mizumoto, A. (2021). New Word Level Checker. https://nwlc.pythonanywhere.com/.
Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins.
Nation, I. S. P. (2017a). *The BNC/COCA Level 6 word family lists (Version 1.0.0)* [dataset] https://people.wgtn.ac.nz/paul.nation.
Nation, I. S. P. (2017b). *The BNC/COCA Level 3 partial word family lists (Version 1.0.0)* [dataset] https://people.wgtn.ac.nz/paul.nation.
Nation, I. S. P. (2022). *Learning vocabulary in another language* (3rd ed.). Cambridge University Press.
Nation, P., & Deweerdt, J. (2001). A defence of simplification. *Prospect, 16*(3), 55–67. https://doi.org/10.26686/wgtn.12560342.v1

Nation, I. S. P., & Waring, R. (2019). *Teaching extensive reading in another language*. Routledge.
NCELP. (2021). *NCELP schemes of work*. UK: National Centre for Excellence for Language Pedagogy. University of York. Retrieved from https://ldpedagogy.org/ncelp-schemes-of-work/ Accessed August 27, 2023.
NCELP. (2022a). *Word patterns*. UK: National Centre for Excellence for Language Pedagogy. University of York. Retrieved from https://resources.ldpedagogy.org/collections/7m01bp08p?locale=en Accessed August 27, 2023.
NCELP. (2022b). *Key stage 4 German scheme of work*. UK: National Centre for Excellence for Language Pedagogy. University of York. Retrieved from https://resources.ncelp.org/concern/resources/d217qs09g?locale=en Accessed August 27, 2023.
Noreillie, A. S., Kestemont, B., Heylen, K., Desmet, P., & Peters, E. (2018). Vocabulary knowledge and listening comprehension at an intermediate level in English and French as foreign languages: An approximate replication study of Stæhr (2009). *ITL - International Journal of Applied Linguistics, 169*(1), 212–231. https://doi.org/10.1075/itl.00013.nor
Nurmukhamedov, U., & Webb, S. (2019). Lexical coverage and profiling. *Language Teaching, 52*(2), 188–200. https://doi.org/10.1017/S0261444819000028
Pinker, S. (1991). Rules of language. *Science, 253*(5019), 530–535. https://doi.org/10.1126/science.1857983
Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.
Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.
Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press.
Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x
Smith, S. (2022). Vocabulary profiler. https://www.eapfoundation.com/vocab/profiler/.
Snoder, P., & Laufer, B. (2022). EFL learners' receptive knowledge of derived words: The case of Swedish adolescents. *Tesol Quarterly, 56*(4), 1242–1265. https://doi.org/10.1002/tesq.3101
Teaching Schools Council. (2016). *Modern foreign languages pedagogy review*. Retrieved from https://pure.york.ac.uk/portal/files/54043904/MFL_Pedagogy_Review_Report_TSC_PUBLISHED_VERSION_Nov_2016_1_.pdf Accessed August 27, 2023.
Tschirner, E., & Möhring, J. (2019). *A frequency dictionary of German: Core vocabulary for learners (2nd ed.)*. Routledge.
Tuggener, D. (2016). Incremental coreference resolution for German (Unpublished doctoral dissertation) Retrieved from https://www.cl.uzh.ch/dam/jcr:b2212d28-6248-47dc-a4e3-04206ff4c6db/tuggener_diss.pdf. Accessed October 2, 2022.
Vold, E. T. (2022). Development of lexical richness among beginning learners of French as a foreign language. *Nordic Journal of Language Teaching and Learning, 10*(2), 182–211. https://doi.org/10.46364/njltl.v10i2.1007
Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. *System, 37*(3), 461–469. https://doi.org/10.1016/j.system.2009.01.004
Webb, S. (2021a). A different perspective on the limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition, 43*(2), 454–461. https://doi.org/10.1017/S0272263121000449
Webb, S. (2021b). Word families and lemmas, not a real dilemma: Investigating lexical units. *Studies in Second Language Acquisition, 43*(5), 973–984. https://doi.org/10.1017/S0272263121000760
West, M. (1953). *A general service list of English words*. Longman, Green & Co.
van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics, 34*(4), 457–479. https://doi.org/10.1093/applin/ams074