Lexical Profiling for Dutch: From Resource to Real-Life Implementation

Jasper Degraeuwe[1], **Natalie Finlayson**[2,3]

[1]LT[3] and MULTIPLES, Ghent University, Belgium
[2]University of York, England
[3]Queen's University Belfast, Northern Ireland

jasper.degraeuwe@ugent.be natalie.finlayson@york.ac.uk

Abstract

Lexical profiling tools use natural language processing (NLP) software to measure the distribution of words in an input text across precompiled word lists. In Intelligent Computer-Assisted Language Learning (ICALL) contexts, foreign/second language (L2) teachers can use the output of lexical profilers to prepare their classes (e.g., by selecting or adapting texts for particular learner groups) or to analyse learner texts. The present study advances the state of the art in lexical profiling by introducing a first-of-its-kind lexical profiling resource for L2 Dutch (dubbed `NLLP-5000`) and demonstrating its use in two concrete ICALL applications. The resource, based on Tiberius and Schoonheim's *A Frequency Dictionary of Dutch* and made publicly available through an open repository, is characterised by the inclusion of inflected forms – obtained through a meticulous manual verification process – and the linguistic annotation of these forms in Universal Dependencies (UD) format. The UD-based annotations in `NLLP-5000` make the resource compatible with lexical profiling tools built on widely used NLP toolkits such as spaCy and Stanza.

`https://github.com/JasperDUGent/NLLP-5000`

1 Introduction

Since the advent of neural networks, deep learning, and (generative) large language models (LLMs), the "Intelligent" in Intelligent Computer-Assisted Language Learning (ICALL) has increasingly come to denote "artificially intelligent". Recent work in the field of ICALL has successfully used AI-driven methods to generate written corrective feedback (Wu et al., 2025), create conversational intelligent tutoring systems (Pérez-Ortiz et al., 2024), tailor texts to the needs of specific learners (Bodnar, 2025), build graded vocabulary lists (Alfter, 2024), and generate pedagogically relevant questions on existing educational materials (Gerald et al., 2026), among other applications and uses.

However, even though the use of AI for educational purposes offers considerable potential and added value, this does not imply that all existing ICALL tools should be replaced by AI or augmented with AI-based components. As argued by Stamper et al. (2024) and Soliar et al. (2026) in the context of intelligent tutoring systems, "one should not disregard previous knowledge and years of experience using expert-generated and traditional data-driven approaches" (Soliar et al., 2026). A similar trend can be observed in the field of corpus linguistics, where LLMs were found to be good at analysing and classifying frequency and collocation data, but to be doing a poor job of generating these data compared to traditional corpus methods (Davies, 2025).

In this paper, we follow this line of thought for the task of lexical profiling, defined by Finlayson et al. (2023) as "the practice of using software to measure the distribution of lexical items in texts across one or more word lists". In the field of ICALL, lexical profiling tools have two main applications: (1) helping teachers prepare foreign/second language (L2) classes and (2) supporting teachers (and researchers) in the analysis and assessment of learner texts.

The contribution of the paper is twofold. First, building on previous knowledge and ex-

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 16–33, November 19–20, 2026, University of Tartu Library

isting resources, we introduce NLLP-5000[1], a first-of-its-kind resource for lexical profiling in Dutch. Second, we integrate NLLP-5000 into MultilingProfiler (Finlayson et al., 2023), a widely used lexical profiling tool, and provide a demonstration of the two above-mentioned applications (class preparation and learner text analysis/assessment).

2 Related Research

2.1 Frequency Dictionaries in SLA

When learning L2 vocabulary, frequency matters (Vilkaitė-Lozdienė and Schmitt, 2019). Although the exact numbers differ slightly across studies and languages, it has been shown that the top 5,000 most frequent word families in a language generally account for approximately 95% of the words in a written text (Green, 2020; Laufer and Ravenhorst-Kalovski, 2010). Consequently, drawing on methods from the domains of corpus linguistics (Sinclair and Renouf, 1988) and lexicography (Frankenberg-Garcia et al., 2021), the field of second language acquisition (SLA) has devoted considerable attention to developing frequency dictionaries that can inform course material developers, researchers, and teachers about which words to prioritise in vocabulary learning activities and test batteries (Nation, 2022).

A good illustration is provided by Brezina and Gablasova (2015), who introduced a new version of the General Service List (West, 1953) for English. Consisting of 2,494 lemmas obtained through a frequency-based selection method, the *new-GSL* covers about 80% of text in a collection of four large corpora (LOB, BNC, BE06, and EnTenTen12, with a combined size of more than 12 billion running words). As a result, L2 teachers and coursebook designers using the *new-GSL* can be reasonably confident that familiarising L2 English learners with its content will enable them to understand most of the words they encounter in running text.

2.2 Lexical Profiling

Apart from informing practitioners during the development of vocabulary learning activities, frequency dictionaries also constitute a valuable resource for lexical profiling tools. Lexical profiling can be used to characterise the vocabulary demands of a text, providing an indication of its lexical load, the potential for vocabulary learning through reading or listening, and the vocabulary knowledge required to understand different text types (Nurmukhamedov and Webb, 2019). Lexical profiling tools do this by comparing the lexical items in a text against one or more word lists representing the expected or targeted vocabulary of a particular learner group or proficiency level, including relevant inflected and derived forms. The proportion of the words in a text that is covered by a given word list is referred to as lexical coverage, and can be used to indicate the extent to which the vocabulary demands of a text are likely to be met by the assumed vocabulary knowledge of a learner group. In pedagogical contexts, such information can help teachers classify, select, and adapt texts for particular learner groups according to their vocabulary level (Dang, 2022).

3 Resource Creation

3.1 Input Data

The first contribution of the present paper is to describe the development of a first-of-its-kind lexical profiling resource for L2 Dutch. As input data to create the resource, *A Frequency Dictionary of Dutch* (Tiberius and Schoonheim, 2014) was used, one of the most extensive efforts to date to present core vocabulary to L2 learners of Dutch. As this resource is part of the *Routledge Frequency Dictionaries* series, it is henceforth referred to as the "Routledge dictionary". The dictionary consists of six separate word lists derived from four corpora[2], each with its own index (a sample from the "Core" list is presented in Figure 1). It is described as containing "the 5,000 top words of Dutch" (Tiberius and Schoonheim, 2014), but because the same headword is allowed to appear in multiple lists[3], the dictionary contains more than 5,000 entries in total. An entry consists of the following five elements:

- **Headword** (boldfaced): usually corresponds to the lemma as lexical unit (see Section 3.2.2 for more details)

[1]For **NL** (Dutch) **L**exical **P**rofiling.

[2]The corpora each represent a different genre: fiction (23 million words), newspaper (167 million), spoken (9 million), and web (91 million). For each corpus, an eponymous "genre list" was created based on a custom procedure (Tiberius and Schoonheim, 2014). The additional "Core" and "General" lists include words occurring frequently across all genres and across at least three of the genres, respectively.

[3]However, the frequency values attached to these duplicate entries are identical across lists. Only the example sentences provided differ depending on the list.

- **Part of speech** (italicised): adjective (`adj`), adverb (`adv`), article (`art`), conjunction (`conj`), interjection (`interj`), noun (`noun`), number (`num`), preposition (`prep`), pronoun (`pron`), or verb (`verb`)
- **English translation** (if the headword has multiple meanings, translation of the most common meaning)
- **Example sentence** (if the headword has multiple meanings, sentence illustrating the most common meaning)
- **Frequency value**: custom dispersion value based on the number of 2,000-word corpus samples in which the headword occurs, normalised across genres

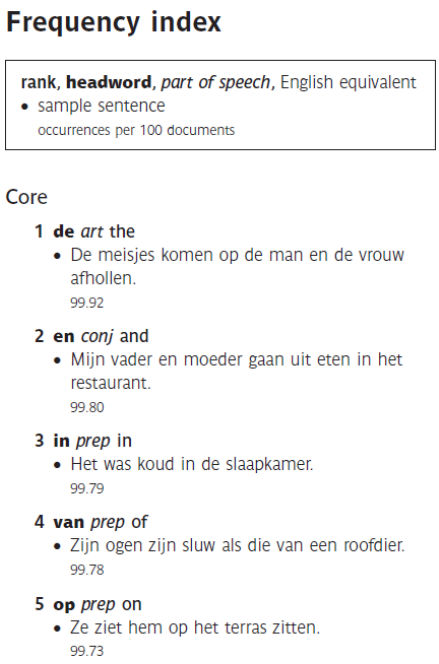


Figure 1: Sample from the "Core" list in *A Frequency Dictionary of Dutch* (Tiberius and Schoonheim, 2014).

3.2 Conversion into Lexical Profiling Resource

To improve consistency and enable seamless integration into existing lexical profiling tools, a series of modifications and additions were made to the input data[4] presented in Section 3.1.

[4]The official TXT file of the dictionary provided by Routledge contains some typographic errors and formatting issues; these were manually fixed. An overview of these fixes is presented in Appendix A. The resulting cleaned file served as input data to create `NLLP-5000`.

3.2.1 Removing Duplicate Entries

Process. Several headwords appeared in more than one of the six lists included in the Routledge dictionary, resulting in duplicate entries for one and the same headword (see Section 3.1). As a first step, these duplicate entries were removed.

Overview of modifications. Of the 5,838 entries included in the Routledge dictionary, 671 duplicate entries were removed. This yielded a provisional resource of 5,167 unique headwords.

3.2.2 Lemmatising Headwords

Process. To ensure consistency, word lists usually use a single type of lexical unit for their headwords. For our resource, we opted for the lemma[5] (see also Section 6 for a brief discussion of lexical units). All unique headwords in the input data which did not correspond to this lexical unit were modified. This task was performed manually by the first author of this paper, who is a native speaker of Dutch and holds a PhD in Linguistics. All decisions were taken based on the *Woordenlijst Nederlandse Taal* as reference work, accessed via Woordenlijst.org[6]. This resource prescribes the official Dutch spelling and provides extensive metadata on all words included in it (e.g., part of speech and inflected forms; see Figure 2 for an illustration). Table 1 illustrates the modifications made during this stage: 24 of the headwords retained after removing duplicates appeared to be inflected forms which needed to be changed into the base form, one entry had more than one part of speech (POS) and needed to be split into separate headwords, and for six entries the POS had to be changed.

After this lemmatisation step, 13 "lemma duplicates" were identified (i.e. word forms appearing as separate entries in the dictionary but belonging to the same lemma; see Table 2). In these cases, the base form was used as the headword and the highest frequency value among the individual word forms was assigned to the lemma. Since the Routledge dictionary includes dispersion-based frequency values and not raw frequencies (see also

[5]That is, a base form combined with a part of speech, including all of its inflected forms. We also consider diminutive forms of nouns (e.g., *dag* ['day'] → *dagje*) to be part of the lemma. Although traditional Dutch grammar treats diminutives as derivations (partly because they can change the noun's gender), they behave like inflections in many other respects (e.g., highly regular and no change in word class).

[6]`https://woordenlijst.org/`

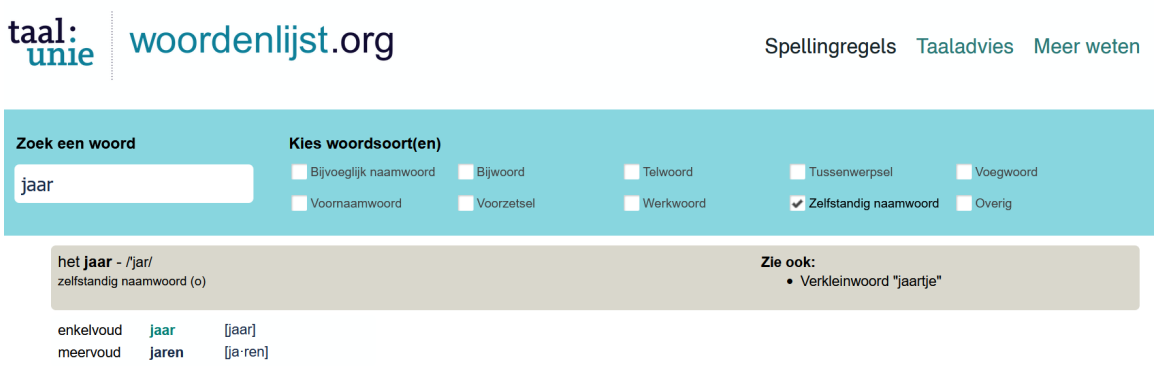


Figure 2: The *Woordenlijst Nederlandse Taal* entry for *jaar* ('year'), accessed via Woordenlijst.org (`https://woordenlijst.org/zoeken/?q=jaar`).

Type	Description	# cases	Applied modification	Instances
Inflected headword	Original headword is an inflected form, not the base form of word	24	Headword changed into base form	All inflected forms in Table 2; *beetje* (diminutive) → *beet*; *inkomsten* (plural) → *inkomst*; *jezelf* (accusative/dative) → *jijzelf*; *kleintje* (diminutive) → *kleine*; *kleren* (plural) → *kleed*; *meerder* as `num` → *meerdere*; *onszelf* (accusative/dative) → *wijzelf*; *praatje* (diminutive) → *praat*; *sommige* → *sommig*; *sprookje* (diminutive) → *sproke*; *werkzaamheden* (plural) → *werkzaamheid*
Multiple POS	Original headword has multiple POS tags assigned to it	1	Headword split into separate headwords	*het* → headword as `art` and as `pron`
Wrong POS	Original headword is tagged with an incorrect POS	6	POS changed	*ander* from `adj` to `num` and `pron`; *dergelijk* from `adj` to `pron`; *anno* from `prep` to `adv`; *alstublieft/alsjeblieft/a.u.b.* from ∅ to `interj`; *cc* from ∅ to `noun`; *o.a.* from ∅ to `adv`

Table 1: Modifications made to the original input data. The list of instances is exhaustive. The ∅ symbol means that no POS was assigned in the Routledge dictionary. Translations are provided in Appendix E.

Section 3.1), the combined frequency value of entries belonging to the same lemma could not be determined: because different word forms of the same lemma may occur in the same samples, summing their values would overestimate the lemma's dispersion. The highest value among the word forms was therefore the best available approximation, albeit a conservative one.

As an illustration, let us consider the forms *laat* ('late'), *later* ('later'), and *laatst* ('latest'), all included as separate entries in the Routledge dictionary. The three forms belong to the same lemma *laat*: *later* is the uninflected comparative form and *laatst* is the uninflected superlative form. *Laat* has a frequency of 30.43, *later* of 35.37, and *laatst* of 44.76. In our resource, the three separate entries are merged into one, namely the base form *laat* as the headword with a frequency value of 44.76 (i.e. the highest among the three individual values).

Overview of modifications. Of the 5,167 entries retained after the first step, 13 "lemma duplicates" were removed (Table 2), 17 entries were modified[7], and two entries were added (one for *het* and one for *ander*). After this step, our provisional resource contained 5,156 headwords in lemma form.

3.2.3 Adding Inflected Forms

Process. The original Routledge dictionary does not provide information on the inflected forms of its entries. However, information on inflected forms is essential for lexical profiling tools (see also Section 2.2). For example, once learners have acquired the basic morphological rules of the L2

[7]This number excludes modifications made for entries which appeared to be lemma duplicates.

frequency	headword_orig	M	O	V	M-O	M-O-V	M-V	O-V	eigennaam	standaardvorm_ENK	standaardvorm_MV	verkleinwoord_ENK	verkleinwoord_MV
1.18	goed		x							goed	goeden	goedje	goedjes
89.77	jaar		x							jaar	jaren	jaartje	jaartjes
73.56	mens	x								mens	mensen	mensje	mensjes
6.71	weer		x							weer	weren	weertje	weertjes
70.73	tijd	x								tijd	tijden	tijdje	tijdjes
7.14	haar									haar	haren	haartje	haartjes
66.05	dag	x								dag	dagen	dagje	dagjes
60.7	plaats						x			plaats	plaatsen	plaatsje	plaatsjes
2.34	enkel	x								enkel	enkelen	enkeltje	enkeltjes

Figure 3: Illustration of the forms automatically populated according to the predefined rules. Incorrect forms were manually corrected (e.g., **goeden* as plural of *goed* ['good'] should become *goederen*) and non-existing forms were replaced by "N/A" (e.g., **weren* as plural of *weer* ['weather'] does not exist).

Base form	**POS**	**Duplicate entries in Routledge dictionary**
laat	`adj`	*later* (comparative), *laatst* (superlative)
ver	`adj`	*verder* (comparative)
goed	`adj`	*best* (superlative)
een	`art`	*ene* (inflected form with *buigings-e*)
goed	`noun`	*goederen* (plural)
kost	`noun`	*kosten* (plural)
medium	`noun`	*media* (plural)
ik	`pron`	*mij/me* (accusative, dative)
hij	`pron`	*hem/'m* (accusative, dative)
zij/ze	`pron`	*hen* (accusative)
jij/je	`pron`	*jou* (accusative, dative)
ikzelf	`pron`	*mijzelf/mezelf* (accusative, dative)

Table 2: Lemma duplicates identified after the lemmatisation process. For translations see Appendix E.

they are studying, they are likely to know the plural form of a noun if they already know its singular form. Consequently, when an L2 teacher uses a lexical profiling tool to prepare instructional materials for such learners, the tool should be able to treat both forms as either known or unknown, even when their frequency values differ substantially.

Therefore, the resource was enriched with the inflected forms listed below. Note that not all headwords have all possible inflections: adjectives such as *huidig* ('current') and *vorig* ('previous') do not have comparative or superlative forms.

- **Adjectives**: positive degree (standard form, inflected form with *buigings-e*[8], and inflected form with *buigings-s*), comparative degree (standard form, inflected form with *buigings-e*, and inflected form with *buigings-s*), and superlative degree (standard form and inflected form with *buigings-e*)
- **Adverbs**: not applicable (in Dutch, adverbs have one invariable form)
- **Articles**: nominative case, accusative case, genitive case, and dative case (the article *een* also has an inflected form with *buigings-e* in nominative, accusative, and dative)
- **Conjunctions**: not applicable (one invariable form)
- **Interjections**: not applicable (one invariable form)
- **Nouns**: standard degree (singular form and plural form) and diminutive degree (singular form and plural form)
- **Numbers**: standard form, inflected form with *buigings-e*, and inflected form with *buigings-n*
- **Prepositions**: not applicable (one invariable form)
- **Pronouns**: nominative case, accusative case, genitive case, and dative case (some pronouns also have inflected forms with *buigings-e* and *buigings-n* in nominative, accusative, and dative)
- **Verbs**: infinitive, present tense forms, past tense forms, imperative, conjunctive, present participle, and past participle

The inflected forms were added using a semi-automatic process. First, a Python script (made available in the `NLLP-5000` repository[9]) was used to automatically populate all forms based on a set of predefined rules, as illustrated in Figure 3.

[8] More information on this and other *buigings* forms can be found in Appendix B.

[9] `https://github.com/JasperD-UGent/NLLP-5000` (`NLLP-5000_compilation`).

Original POS tag	UD tag(s)	Mapping type	Mapping details
adj (adjective)	ADJ	1-to-1	Mapping to fully equivalent UD ADJ tag.
adv (adverb)	ADV ADP PART	Manual	Regular adverbs were mapped to the ADV tag; verbal particles were mapped to the ADP tag; particles encoding negation (e.g., *niet* ['not']) were mapped to the PART tag.
art (article)	DET	1-to-1	In UD, articles are considered determiners (i.e. words that modify nouns or noun phrases and express the reference of the noun phrase in context). Therefore, the original art tag was mapped to the UD DET tag.
conj (conjunction)	CCONJ SCONJ	Manual	In UD, conjunctions are divided into coordinating conjunctions (CCONJ; e.g., *of* ['or']) and subordinating conjunctions (SCONJ; e.g., *hoewel* ['although']). All conjunctions in the dictionary were mapped to their corresponding type.
interj (interjection)	INTJ	1-to-1	Mapping to fully equivalent UD INTJ tag.
noun (noun)	NOUN PROPN	Manual	In UD, nouns are divided into common nouns (NOUN; e.g., *aanval* ['attack']) and proper nouns (PROPN; e.g., *Bijbel* ['Bible']). All nouns in the dictionary were mapped to their corresponding type.
num (number)	NUM	1-to-1	Mapping to fully equivalent UD NUM tag.
prep (preposition)	ADP	1-to-1	In UD, prepositions form part of the broader adposition category, which also encompasses circum- and postpositions. Therefore, the original prep tag was mapped to the UD ADP tag.
pron (pronoun)	PRON	1-to-1	Mapping to fully equivalent UD PRON tag.
verb (verb)	AUX VERB	Manual	In UD, verbs are divided into auxiliary verbs (AUX; e.g., *kunnen* ['can']) and main verbs (VERB; e.g., *creëren* ['to create']). Note that some verbs can be used as both types (e.g., *gaan* ['to go']). All verbs in the dictionary were mapped to their corresponding type(s).

Table 3: Overview of mapping from POS tags used in original dictionary to UD tags.

Second, all these forms were manually checked by the first author of the paper.

Overview of additions. In this step, 40,169 inflected forms were added to the 5,156 headwords obtained in the previous step. Manual verification ensured a high degree of data quality, while the exhaustive inventory of inflected forms gives the resource its unique, first-of-its-kind character.

3.2.4 Enriching Linguistic Information

Process. To allow for compatibility with lexical profiling tools offering more advanced features (e.g., removing certain morphological inflections), all inflected word forms added in the previous step were tagged with POS tags and morphological features in the Universal Dependencies (UD) format (De Marneffe et al., 2021). UD is a framework for cross-linguistically consistent annotation of human language at a grammatical level (i.e. labelling of parts of speech, morphological features, and syntactic dependencies). Covering more than 150 languages and used by popular natural language processing (NLP) toolkits such as spaCy[10] and Stanza (Qi et al., 2020), UD has become one of the most widely used frameworks for morphosyntactic analysis in the field of computational linguistics.

Table 3 shows how the POS tags used in the original dictionary were mapped to UD tags. One-to-one mappings were performed automatically; manual mappings were performed by the first author. More information on the definition of the morphological features is presented in Appendix C. The Python scripts used to add the UD POS tags and morphological features to the resource are publicly available in the above-mentioned GitHub repository.

Overview of additions. In this step, the 40,169 inflected forms (belonging to 5,156 unique headwords) from the previous step were enriched with UD POS tags and morphological features, allowing for the seamless integration of the resource into existing lexical profiling tools. This integration process will be described in Section 4 and demonstrated in Section 5.

[10] https://spacy.io/

3523-1	3,523	4	verlichten	verb	verlichten	VERB	VerbForm=Inf
3523-2	3,523	4	verlichten	verb	verlicht	VERB	Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin
3523-3	3,523	4	verlichten	verb	verlicht	VERB	Mood=Ind|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin
3523-4	3,523	4	verlichten	verb	verlicht	VERB	Mood=Ind|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin
3523-5	3,523	4	verlichten	verb	verlicht	VERB	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin
3523-6	3,523	4	verlichten	verb	verlichten	VERB	Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin
3523-7	3,523	4	verlichten	verb	verlichten	VERB	Mood=Ind|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin
3523-8	3,523	4	verlichten	verb	verlichten	VERB	Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin
3523-9	3,523	4	verlichten	verb	verlichtte	VERB	Mood=Ind|Number=Sing|Person=1|Tense=Past|VerbForm=Fin
3523-10	3,523	4	verlichten	verb	verlichtte	VERB	Mood=Ind|Number=Sing|Person=2|Tense=Past|VerbForm=Fin
3523-11	3,523	4	verlichten	verb	verlichtte	VERB	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin
3523-12	3,523	4	verlichten	verb	verlichtten	VERB	Mood=Ind|Number=Plur|Person=1|Tense=Past|VerbForm=Fin
3523-13	3,523	4	verlichten	verb	verlichtten	VERB	Mood=Ind|Number=Plur|Person=2|Tense=Past|VerbForm=Fin
3523-14	3,523	4	verlichten	verb	verlichtten	VERB	Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin
3523-15	3,523	4	verlichten	verb	verlicht	VERB	Mood=Imp|VerbForm=Fin
3523-16	3,523	4	verlichten	verb	verlichte	VERB	Mood=Sub|VerbForm=Fin
3523-17	3,523	4	verlichten	verb	verlichtend|verlichtende	VERB	Tense=Pres|VerbForm=Part
3523-18	3,523	4	verlichten	verb	verlicht	VERB	Tense=Past|VerbForm=Part
3649-1	3,649	4	opinie	noun	opinie	NOUN	Gender=Fem|Number=Sing
3649-2	3,649	4	opinie	noun	opinies	NOUN	Gender=Fem|Number=Plur
3649-3	3,649	4	opinie	noun	opinietje	NOUN	Degree=Dim|Gender=Fem|Number=Sing
3649-4	3,649	4	opinie	noun	opinietjes	NOUN	Degree=Dim|Gender=Fem|Number=Plur
3683-1	3,683	4	badkamer	noun	badkamer	NOUN	Gender=Fem,Masc|Number=Sing
3683-2	3,683	4	badkamer	noun	badkamers	NOUN	Gender=Fem,Masc|Number=Plur
3683-3	3,683	4	badkamer	noun	badkamertje	NOUN	Degree=Dim|Gender=Fem,Masc|Number=Sing
3683-4	3,683	4	badkamer	noun	badkamertjes	NOUN	Degree=Dim|Gender=Fem,Masc|Number=Plur
4030-1	4,030	5	tevergeefs	adj	tevergeefs	ADJ	Degree=Pos
4030-2	4,030	5	tevergeefs	adj	tevergeefse	ADJ	Degree=Pos
4030-3	4,030	5	tevergeefs	adj	tevergeefs	ADJ	Degree=Pos
4036-1	4,036	5	erom	adv	erom	ADV	N/A

Created with Datawrapper

Figure 4: Sample of NLLP-5000. The visualisation was created with Datawrapper.

3.2.5 Converting Frequency Values

Process. Finally, the 5,156 unique headwords were ranked from highest frequency to lowest. Only the top 5,000 items were retained, meaning that the 156 items with the lowest frequency were removed from the resource. For the 5,000 selected items, we determined two new frequency-related values:

- **Frequency rank**: position of the headword in the ranked list (1 for headword with highest frequency; 5,000 for headword with lowest frequency)
- **Frequency band**: when dividing the list into five equal parts (1–1,000 | 1,001–2,000 | 2,001–3,000 | 3,001–4,000 | 4,001–5,000), part to which the headword belongs

Overview of modifications. In this step, the resource was trimmed down from 5,156 to 5,000 headwords and the original dispersion values were replaced by frequency rank and frequency band values. The set of 5,000 headwords in lemma form accounts for a total of 38,811 inflected forms.

3.3 Final Resource

Figure 4 shows a sample of the final version of our lexical profiling resource, dubbed NLLP-5000. The main resource statistics are presented in Table 4. The full resource is made available in the GitHub repository[11]. Note that the English translation and the example sentences (see Section 3.1) are not included in the resource, as the

[11] https://github.com/JasperD-UGent/NLLP-5000 (NLLP-5000_data).

Band	adj	adv	art	conj	interj	noun	num	prep	pron	verb	Inflections
1–1,000	156	123	3	18	4	359	30	38	50	219	7,397
1,001–2,000	177	67	0	7	6	480	16	5	5	237	7,625
2,001–3,000	167	54	0	4	4	507	4	6	3	251	7,759
3,001–4,000	172	36	0	4	3	524	6	2	3	250	7,780
4,001–5,000	161	61	0	5	7	461	7	1	4	293	8,250
Total	833	341	3	38	24	2,331	63	52	65	1,250	38,811

Table 4: Main statistics for `NLLP-5000`: number of headwords per POS and number of inflections, reported per frequency band and in total.

Inflected form	UD POS	UD features	FreeLing tag
komt ('comes')	VERB	Mood=Ind | Number=Sing | Person=3 | Tense=Pres | VerbForm=Fin	`VVIPS3`
kom ('come')	VERB	Mood=Imp | VerbForm=Fin	`VVM000`
gekomen ('come')	VERB	Tense=Past | VerbForm=Part	`VVPS00`

Table 5: Examples of the conversion from UD to FreeLing, three inflected forms of the headword *komen* (`verb`).

Routledge licence agreement prohibits redistributing these data.

4 Implementation in Existing Tools

Although lexical profiling tools serve the same purpose, they often employ different procedures for processing input texts and require their source lists to conform to tool-specific formats. As a result, integrating existing word lists into these tools can be cumbersome. When building the `NLLP-5000` resource, we anticipated this issue by enriching all word forms in the resource with detailed linguistic information (POS tag and morphological features) in the standardised UD format (see Section 3.2.4). With simple rule-based scripts, the UD-based entries from our resource can easily be converted into other formats with little or no loss of information.

In the present study, we illustrate this conversion process for the **MultilingProfiler** tool (Finlayson et al., 2023), which expects source lists in a FreeLing-based format (Padró and Stanilovsky, 2012). The FreeLing tags constitute essential information for the functionalities involving morphological inflections. MultilingProfiler's "Remove past tense forms" feature, for example, can prove to be very useful for teachers whose learners have yet to learn this verbal inflection and who therefore want the profiler to mark all past tense forms as unknown. To illustrate the format conversion, let us consider the POS category of verbs, for which the following FreeLing feature-value pairs[12] were required:

- **POS letter**: always `V`
- **Verb type**: `V` for lexical ("full") verbs, `A` for regular auxiliary verbs, `M` for modal verbs
- **Mood**: `C` for conjunctive/subjunctive, `I` for indicative, `M` for imperative, `N` for infinitive, `P` for participle, `0` if not applicable
- **Tense**: `P` for present, `S` for past, `0` if not applicable
- **Number**: `P` for plural, `S` for singular, `0` if not applicable
- **Person**: `1` for first person, `2` for second person, `3` for third person, `0` if not applicable

Using a rule-based script (Appendix D), the appropriate FreeLing tag could be assigned automatically to all verb entries (see Table 5 for examples).

[12]The final FreeLing tag is obtained by concatenating the values of all features into a single string (e.g., `VVIPS1` for a lexical verb in the indicative mood, present tense, and first-person singular).



Figure 5: Demonstration of preparing a text for a B2-level L2 Dutch class using MultilingProfiler and `NLLP-5000` (profiler output).

The same procedure was applied to the remaining POS categories.

5 ICALL Applications

5.1 Class Preparation

Description. Consider a teacher preparing a B2-level class on the topic of "nature". The school curriculum assumes that learners at this level are already familiar with the 5,000 most frequent lemmas in Dutch (i.e. the lemmas included in `NLLP-5000`). When selecting a text for classroom use, the teacher may therefore wish to identify words outside this assumed vocabulary. These words can then be addressed in a preparatory vocabulary activity or, in case of a reading-focused lesson, provided in a glossary.

To illustrate this process, let us suppose the teacher found an interesting news article online (see Appendix G for the full text). The teacher enters the article into MultilingProfiler, which then profiles the text against the `NLLP-5000` resource and highlights all words not covered by this list (Figure 5). The tool also provides summary statistics, including the percentage of running words covered by the list (Figure 6), and allows teachers to deselect certain inflections (see example of removing past tense forms discussed in Section 4). These outputs allow the teacher to quickly assess the lexical demands of the text and identify vocabulary that may require additional attention.

Discussion. In this application, it is assumed that the target learners are familiar with the 5,000 lemmas included in the `NLLP-5000` resource. Naturally, this assumption will only hold if learning these 5,000 lemmas forms part of the pre-B2 phase of the study curriculum. In language learning curricula taking learner-oriented corpus-based vocabulary lists (e.g., Kilgarriff et al., 2014) as their starting point, different assumptions will need to be made. To this end, MultilingProfiler offers the possibility to upload a custom list to be used as reference point for the profiling.

5.2 Learner Text Analysis and Assessment

Description. Alternatively, teachers can use `NLLP-5000` in a lexical profiling tool to analyse vocabulary use in learner-produced text (the results of which could then be used as a source of evidence in assessment). Consider the authentic learner text presented in Appendix F. By profil-

Global coverage

Total words covered by all lists (including multiword units)	446	82.1%

Word statistics (Tokens; each word counted separately. Figures show text coverage.)	count	% coverage
Total words from your chosen list	446	82.1%
Total words from your extended list	0	0.0%
Total cognates from outside your chosen list	0	0.0%
Total words from all lists (chosen list + extended list + cognates)	446	82.1%
Total words in the text	543	
Type-token ratio	0.5	

Figure 6: Demonstration of preparing a text for a B2-level L2 Dutch class using MultilingProfiler and NLLP-5000 (statistics).



Figure 7: Demonstration of analysing learner text using MultilingProfiler and NLLP-5000 (profiler output).

ing this text in MultilingProfiler, teachers can obtain an indication of how advanced the learner's vocabulary usage is. If learners are expected to know the 4,000 most frequent lemmas, for example, the teacher can select this threshold in the "Band" dropdown (Figure 7) and examine how many words in the text belong to frequency bands beyond this threshold. This information is also available in the statistics file, which can be downloaded in several different file formats and enables users to sort every word in the text into the selected frequency bands.

Discussion. This type of analysis should not be interpreted as an error analysis. While it is true that non-existing word forms will be classified as off-list, some learner errors involve incorrect combinations or uses of otherwise valid words and will therefore not be highlighted. This distinction is particularly important when lexical profiling is used to support assessment. A case in point is the phrase "met de lokale bevolking rijst" in Figure 7: *rijst* is an existing Dutch word meaning 'rice', but in this context the learner intended *reist* ('travels'). The lexical profiler, however, treats *rijst* as a valid lexical item even though its use in the sentence is erroneous.

Additionally, since NLLP-5000 does not distinguish word senses, the profiling results should not be interpreted as conclusive evidence of the learner's level of lexical proficiency. On the one hand, a learner showing productive knowledge of a polysemous lemma does not imply that this learner has acquired knowledge of all senses of this lemma. Especially for lemmas whose senses are substantially different in terms of frequency

and difficulty (e.g., *slot* can mean both 'lock' and 'castle'), this is an important observation to bear in mind. On the other hand, the use of relatively infrequent items alone does not necessarily mean that the learner has advanced lexical proficiency.

6 Discussion and Future Work

By introducing `NLLP-5000`, the present study advances the state of the art for lexical profiling in (L2) Dutch. Compared to the few existing Dutch word frequency lists designed for lexical profiling (e.g., Van Parys et al., 2025), our resource is distinguished by the inclusion of inflected forms (obtained through a meticulous manual verification process) and the linguistic annotation of these forms in Universal Dependencies (UD) format.

The implementation of `NLLP-5000` in an existing lexical profiling tool was demonstrated through an end-to-end workflow: from the conversion of UD-based labels into tool-specific tags (Section 4) to the illustration of concrete applications in ICALL contexts (Section 5). In both ICALL applications demonstrated, the focus was on L2 teachers as the primary stakeholders. This focus was deliberate: because recent research has shown that teachers are underrepresented in educational technology (EdTech) research (Gaudeau et al., 2026), we sought to contribute to a better representation of teachers in this field.

Future work should first of all address the limitations discussed in Section 8, such as the lack of word sense disambiguation and the limited size of the resource. In addition to this, extending the resource with the dimension of word families (Bauer and Nation, 1993) also constitutes an avenue worth exploring. Though not entirely undisputed as a measure for determining vocabulary size (Laufer, 2021), word family forms (e.g., *to use–usable*) have been argued to be the most transparent and efficient lexical unit from a pedagogical point of view, at least for more advanced learners (Webb, 2021). Conducting research on a word family resource for Dutch would constitute a valuable contribution to a field that is, apart from a few exceptions such as Volodina et al. (2024b) for Swedish, dominated by English both in terms of research and available resources.

7 Conclusion

This study built on previous knowledge and existing resources to develop a novel resource for lexical profiling in Dutch. Drawing on an extensive body of research emphasising the vital role of frequency in L2 vocabulary learning, we used *A Frequency Dictionary of Dutch* (Tiberius and Schoonheim, 2014) as input data. Several steps were undertaken to transform these input data into the novel `NLLP-5000` resource (**NL** **L**exical **P**rofiling, comprising 5,000 headwords), which is made publicly available through a GitHub repository. Duplicate entries were removed, all headwords were lemmatised, inflected forms were added and labelled according to the Universal Dependencies (UD) format, and the original dispersion values were converted into frequency ranks and bands.

The UD-based linguistic annotations in `NLLP-5000` make the resource compatible with lexical profiling tools built on widely used NLP toolkits such as spaCy and Stanza. The list is versatile in that it can be easily incorporated into tools that require a different format. We demonstrated this conversion process in Section 4 for MultilingProfiler, alongside two applications of the resource in ICALL contexts: class preparation and learner text analysis/assessment (Section 5).

Finally, although the present paper has mainly focused on the task of lexical profiling, the `NLLP-5000` resource can be integrated more broadly into ICALL environments. Whether in ICALL tools specifically dedicated to vocabulary learning (e.g., Goethals, 2018; for Spanish) or in full-fledged intelligent tutoring systems (e.g., Amaral and Meurers, 2011; for Portuguese), vocabulary lists constitute a valuable source of input data for generating well-informed vocabulary learning exercises tailored to the specific needs of both teachers and (groups of) learners.

8 Limitations

Word senses. In the current version of the resource, words with multiple senses appear under the same headword (e.g., *koper* as 'buyer' and as 'copper'). Extended work could introduce "word sense awareness" into the resource by using word sense disambiguation (WSD) methods for L2 learning purposes (Degraeuwe and Goethals, 2022).

Receptive versus productive knowledge. The class preparation application (Section 5.1) and the analysis of learner text (Section 5.2) address different aspects of vocabulary knowledge. Whereas

the former focuses on the vocabulary learners should recognise in educational materials (i.e. *receptive* knowledge), the latter aims to provide insights into their *productive* knowledge. These two aspects of vocabulary knowledge do not necessarily follow the same frequency thresholds. Following the work of Volodina et al. (2024a) on the Swedish Second Language Profile, future work should explore creating a suite of resources that clearly distinguish receptive from productive knowledge (e.g., with separate frequency indices).

Resource size. `NLLP-5000` is limited to the top 5,000 most frequent lemmas in Dutch. For beginner and intermediate learners, this range is large enough for the lexical profiler to return relevant output. For teachers working with advanced learners, however, the resource will be less useful. As advanced learners have likely mastered most or even all of the 5,000 lemmas included in the list, an additional set of frequency bands is required to profile texts in a meaningful way (e.g., up to the top 15,000 most frequent lemmas).

Validation. The applications presented in Section 5 constitute an *illustration* of how `NLLP-5000` can be used in an ICALL context; they should not be interpreted as *validated* pedagogical practices. To validate the resource, future work should study if the frequency ranks and bands correspond to learners' vocabulary knowledge and/or instructional needs.

9 Ethical Considerations

Role of paper author as annotator. The first author of the paper was responsible for (1) lemmatising the original Routledge input data, (2) verifying the automatically populated inflected forms, and (3) labelling the inflected forms according to the UD format. All decisions taken during these tasks (e.g., assigning base forms to POS categories) were taken based on the official prescriptive resource for Dutch, Woordenlijst.org (described in detail in Section 3.2.2). All other aspects of the tasks (e.g., checking if automatically populated forms were correct) did not entail any form of subjective judgement. We therefore conclude that the author's involvement in the methodology does not compromise the validity of the newly created resource.

Learner data. The writing assignment submission used to demonstrate the second ICALL application (Section 5.2) was fully anonymous. No personal data of the student were processed or stored in any way. The assignment itself also does not include any data that could potentially identify the student.

Acknowledgments

We wish to express our sincere gratitude to the CVO Groeipunt student who was willing to share the writing assignment submission presented in Appendix F and to the anonymous reviewers for their constructive and very valuable feedback.

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: An AI tool was used for flagging language errors and spotting inconsistencies (e.g., using the same terms for the same concepts throughout the paper) as a final check prior to submitting the camera-ready version of the paper.

References

David Alfter. 2024. Out-of-the-Box Graded Vocabulary Lists with Generative Language Models: Fact or Fiction? In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 1–19, Rennes, France. LiU Electronic Press.

Luiz A. Amaral and Detmar Meurers. 2011. On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL*, 23(1):4–24.

Laurie Bauer and I.S.P. Nation. 1993. Word Families. *International Journal of Lexicography*, 6(4):253–279.

Stephen Bodnar. 2025. A prototype authoring tool for editing authentic texts using LLMs to increase support for contextualised L2 grammar practice. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 63–71, Tallinn, Estonia. University of Tartu Library.

Vaclav Brezina and Dana Gablasova. 2015. Is There a Core General Vocabulary? Introducing the New General Service List. *Applied Linguistics*, 36(1):1–22.

Thi Ngoc Yen Dang. 2022. Using VocabProfilers to select texts for extensive reading activities. In *Teaching English with Corpora*, pages 69–73. Routledge, London.

Mark Davies. 2025. CORPORA AND AI / LLMs: Overview. Online report, English-Corpora.org.

Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jasper Degraeuwe and Patrick Goethals. 2022. Interactive Word Sense Disambiguation in Foreign Language Learning. In *Proceedings of the 11th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL 2022)*, pages 46–54.

Natalie Finlayson, Emma Marsden, and Laurence Anthony. 2023. Introducing MultilingProfiler: An adaptable tool for analysing the vocabulary in French, German, and Spanish texts. *System*, 118:103122.

Ana Frankenberg-Garcia, Geraint Paul Rees, and Robert Lew. 2021. Slipping Through the Cracks in e-Lexicography. *International Journal of Lexicography*, 34(2):206–234.

Gabrielle Gaudeau, Aoife O'Driscoll, Jasper Degraeuwe, Andrew Caines, Donya Rooein, and Zeerak Talat. 2026. Incentives Of EdTech: A Systematic Review Of EduNLP Research. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 715–750, San Diego, California, USA. Association for Computational Linguistics.

Thomas Gerald, Sahar Ghannay, Julie Lascar, Paul Lerner, and Anne Vilnat. 2026. Can Multimodal LLMs Generate Pedagogical Questions? pages 5506–5515, Palma, Mallorca, Spain.

Patrick Goethals. 2018. Customizing vocabulary learning for advanced learners of Spanish. In *Technological innovation for specialized linguistic domains: languages for digital lives and cultures, proceedings of TISLID'18*, pages 229–240, Gent, Belgium. Éditions Universitaires Européennes.

Clarence Green. 2020. Extensive reading and viewing as input for academic vocabulary: A large-scale vocabulary profile coverage study of students' reading and writing across multiple secondary school subjects. *Lingua*, 239:102838.

Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48(1):121–163.

Batia Laufer. 2021. Lemmas, flemmas, word families, and common sense. *Studies in Second Language Acquisition*, 43(5):965–968.

Batia Laufer and Geke C. Ravenhorst-Kalovski. 2010. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1):15–30.

I.S.P. Nation. 2022. *Learning Vocabulary in Another Language*, 3 edition. Cambridge University Press.

Ulugbek Nurmukhamedov and Stuart Webb. 2019. Lexical coverage and profiling. *Language Teaching*, 52(02):188–200.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.

Juan Antonio Pérez-Ortiz, Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Felipe Sánchez-Martínez, Roman Chernysh, Gabriel Mora-Rodríguez, and Lev Berezhnoy. 2024. A Conversational Intelligent Tutoring System for Improving English Proficiency of Non-Native Speakers via Debriefing of Online Meeting Transcriptions. pages 187–198.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

John McHardy Sinclair and Antoinette Renouf. 1988. A lexical syllabus for language learning. In Ronald Carter and Michael Mccarthy, editors, *Vocabulary and Language Teaching*. Routledge.

Mariia Soliar, Leona Colling, Stephen Bodnar, and Detmar Meurers. 2026. Using Interaction Log Data to Evaluate and Improve Feedback Accuracy in an Intelligent Language Tutoring System. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 190–207, San Diego, California, USA. Association for Computational Linguistics.

John Stamper, Ruiwei Xiao, and Xinying Hou. 2024. Enhancing LLM-Based Feedback: Insights from Intelligent Tutoring Systems and the Learning Sciences. Version Number: 2.

Carole Tiberius and Tanneke Schoonheim. 2014. *A frequency dictionary of Dutch: Core vocabulary for learners*. Routledge frequency dictionaries. Taylor & Francis Routledge, London.

Amaury Van Parys, Vanessa De Wilde, Lieve Macken, and Maribel Montero Perez. 2025. LexPro: A plurilingual lexical profiling tool to assist teachers and researchers in analysing vocabulary of L2 input. *Language Teaching Research*, pages 1–25.

Laura Vilkaitė-Lozdienė and Norbert Schmitt. 2019. Frequency as a Guide for Vocabulary Usefulness. In Stuart Webb, editor, *The Routledge Handbook of Vocabulary Studies*, 1 edition, pages 81–96. Routledge.

Elena Volodina, David Alfter, and Therese Lindström Tiedemann. 2024a. Profiles for Swedish as a Second Language: Lexis, Grammar, Morphology. pages 10–19.

Elena Volodina, Yousuf Ali Mohammed, and Therese Lindström Tiedemann. 2024b. Swedish word family resource: Construction, applicability, strengths and first experiments. *ITL - International Journal of Applied Linguistics*, 175(1):127–161.

Stuart Webb. 2021. Word families and lemmas, not a real dilemma: Investigating lexical units. *Studies in Second Language Acquisition*, 43(5):973–984.

Michael West. 1953. *A General Service List of English Words*. Longmans, Green & Co., London.

Jiahui Wu, Jianwei Li, Zigang Ge, Mingrui Xu, Li Lin, and Ru Zhang. 2025. Effectiveness of Generative AI in Automated Written Corrective Feedback With Prompting. *Journal of Educational Computing Research*, 63(6):1493–1527.

A Pre-processing of Original Routledge File

As explained in Sections 3.1 and 3.2, *A Frequency Dictionary of Dutch* (Tiberius and Schoonheim, 2014) was used as input data to create the `NLLP-5000` resource. The dictionary was accessed through the "Frequency-index.txt" file, provided by Routledge as an official supplement to the dictionary. However, upon processing the file, a series of small (formatting) issues were found. These issues, listed below, were manually fixed by the authors of this paper.

- **Line 7**: formatting error, no tab before second example sentence (missing tab was added and one of the trailing tabs was removed)
- **Line 1013**: typo in frequency of *kleden* (". 3.80" instead of "3.80")
- **Line 1344**: formatting error, only 10 elements in line instead of 11 (missing tab was added before "2.47")
- **Line 2471**: typo in frequency of *set* (". 3.98" instead of "3.98")
- **Line 3658**: typo in frequency of *ontwerp* (". 3.49" instead of "3.49")
- **Line 4157**: typo in frequency of *zichtbaar* (". 5.74" instead of "5.74")
- **Line 4865**: incomplete POS ("2) het" instead of "2) noun, het")
- **Lines 5202-5203**: formatting error (both lines belong to *vet*, so they were merged)
- **Lines 5405-5406**: formatting error (both lines belong to *rot*, so they were merged)

B Description of *buigings*-forms

- *Buigings-e* (adjectives): In some cases, an *e* is added to adjectives when they are used attributively (e.g., with plural nouns). For more detailed information, please refer to the official Taaladvies page.
- *Buigings-s* (adjectives): Inflected form needed when the adjective is combined with *iets* ('something').
- *Buigings-e* (articles): The article *een* ('a, an') has the inflected form *ene* in nominative, accusative, and dative.
- *Buigings-e* (numbers): In some cases, an *e* is added to numerals (e.g., *meest–meeste* ['most']), for example when the numeral is used attributively.
- *Buigings-n* (numbers): In some cases, *en* is added to numerals (e.g., *tweeën*), for example in the phrase *met z'n* ... (please refer to the Taaladvies page for more details).
- *Buigings-e* (pronouns): In some cases, an *e* is added to pronouns (e.g., *ieder–iedere* ['each']), for example when the pronoun is used attributively.
- *Buigings-n* (pronouns): In some cases, *(e)n* is added to pronouns (e.g., *al–allen* ['all']), for example when the pronoun is used to refer to a group of people.

C Definition of Morphological Features

The complete overview of the UD morphological features assigned to the inflected forms in `NLLP-5000` is presented in Table 6.

Original POS tag	UD feature	Applicable values for feature	Example(s)
adj (adjective)	Degree	Pos (positive), Cmp (comparative), Sup (superlative)	*kleiner* ('smaller'): Degree=Cmp
adv (adverb)	PartType (particle type)	Neg (negation particle), Vbp (verbal particle)	*niet* ('not'): PartType=Neg | Polarity=Neg *nooit* ('never'): PronType=Neg *weer* ('re-'): PartType=Vbp
	Polarity	Neg (negative)	
	PronType (pronominal type)	Neg	
art (article)	Case	Nom (nominative), Acc (accusative), Gen (genitive), Dat (dative)	*de* ('the'): Case=Nom | Definite=Def | PronType=Art *ener* ('of a, of an'): Case=Gen | Definite=Ind | PronType=Art
	Definite	Def (definite), Ind (indefinite)	
conj (conjunction)	N/A	N/A	N/A
interj (interjection)	N/A	N/A	N/A
noun (noun)	Degree	Dim (diminutive)	*aanbieding* ('offer'): Gender=Fem | Number=Sing *touwtjes* ('little ropes'): Degree=Dim | Gender=Neut | Number=Plur
	Gender	Fem (feminine), Masc (masculine), Neut (neuter)	
	Number	Sing (singular), Plur (plural)	
num (number)	Definite	Def (definite), Ind (indefinite)	*acht* ('eight'): Definite=Def | NumType=Card *ander* ('other'): Definite=Ind | NumType=Ord
	NumType (numeral type)	Card (cardinal), Ord (ordinal)	
prep (preposition)	AdpType (adposition type)	Circ (circumposition), Post (postposition), Prep (preposition)	*achter* ('behind'): AdpType=Prep
pron (pronoun)	Case	Nom, Acc, Gen, Dat	*aller* ('of all'): Case=Gen | Number=Sing | PronType=Ind *deze* ('this'): Case=Nom | Number=Sing | PronType=Dem
	Gender	Fem, Masc, Neut	
	Number	Sing, Plur	
	Person	1, 2, 3	
	Poss (possessive)	Yes	
	PronType	Dem (demonstrative), Ind (indefinite), Int (interrogative), Prs (personal), Rcp (reciprocal), Rel (relative)	
	Reflex	Yes	
verb (verb)	Mood	Imp (imperative), Ind (indicative), Sub (subjunctive)	*buigt* ('bends'): Mood=Ind | Number=Sing | Person=3 | Tense=Pres | VerbForm=Fin *geparkeerd* ('parked'): Tense=Past | VerbForm=Part
	Number	Sing, Plur	
	Person	1, 2, 3	
	Tense	Past, Pres (present)	
	VerbForm	Fin (finite), Inf (infinitive), Part (participle)	
	VerbType	Cop (copula), Mod (modal)	

Table 6: UD morphological features assigned to the inflected forms in NLLP-5000. "N/A" stands for not applicable.

D Script for Converting Verb Entries from UD to FreeLing Format

The script used to automatically convert verb entries in `NLLP-5000` from UD format to FreeLing format is included as supplementary data in the GitHub repository: https://github.com/JasperD-UGent/NLLP-5000/tree/main/supplementaryMaterials/NLP4CALL2026.

E Translation Glossaries

Translation glossary for Table 1

- *beetje* = 'bit'
- *beet* = 'bite'
- *inkomsten* = 'income, revenue'
- *inkomst* = 'entry'
- *jezelf* = 'yourself'
- *jijzelf* = 'yourself'
- *kleintje* = 'small/little one; small thing'
- *kleine* = 'little one'
- *kleren* = 'clothes'
- *kleed* = 'carpet; robe'
- *meerdere* = 'multiple'
- *onszelf* = 'ourselves'
- *wijzelf* = 'ourselves'
- *praatje* = 'chat'
- *praat* = 'talk'
- *sommig* = 'some'
- *sprookje* = 'fairy tale'
- *sproke* = 'tale'
- *werkzaamheden* = 'work, operations'
- *werkzaamheid* = 'activity'
- *het* = 'the; it'
- *ander* = 'other'
- *dergelijk* = 'similar'
- *anno* = 'in the year'
- *alstublieft/alsjeblieft/a.u.b.* = 'please'
- *cc* = 'carbon copy; cubic centimetre'
- *o.a.* = 'among other things, among others'

Translation glossary for Table 2

- *laat* = 'late'
- *later* = 'later'
- *laatst* = 'latest'
- *ver* = 'far'
- *verder* = 'further'
- *goed* = 'good'
- *best* = 'best'
- *een* = 'a, an'
- *ene* = 'a, an'
- *goed* = 'good, property'
- *goederen* = 'goods'
- *kost* = 'cost; living'
- *kosten* = 'cost(s), expense(s)'
- *medium* = 'medium'
- *media* = 'media'
- *ik* = 'I'
- *mij/me* = 'me'
- *hij* = 'he'
- *hem/'m* = 'him'
- *zij/ze* = 'she; they'
- *hen* = 'them'
- *jij/je* = 'you'
- *jou* = 'you'
- *ikzelf* = 'I myself'
- *mijzelf/mezelf* = 'myself'

F Input Text Learner Text Analysis and Assessment Demo

- Source: anonymous writing assignment submission from an L2 Dutch Level 4 course organised at CVO Groeipunt (Flemish adult education centre)
- Date: 28 February 2026

Voor mijn volgende reis naar Sri Lanka ben ik goed voorbereid omdat ik op een forum een artikel over de verschillende vervoermiddelen gelezen heb. Ik heb gelezen dat je met de bus, de trein, per tuktuk of taxi of met een privéchauffeur kunt reizen. Op het forum hebben toeristen verschillende manieren beschreven, met welke zij in Sri Lanka gereisd zijn. Er is een groot verschil tussen de vervoermiddelen, sommige zijn goedkoop maar misschien niet zo comfortabel en andere zijn een beetje duur maar ook heel flexibel. Blijkbaar is het het tofste om met de trein te reizen omdat het een avontuur is, je snel door het land komt en, het belangrijkste, dat je van een heel mooi uitzicht kunt genieten. Op de trein kan je tussen drie klassen kiezen maar de tweede is blijkbaar de beste manier om te reizen omdat je een stoel krijgt maar ook samen met de lokale bevolking rijst. Met de bus kan je ook reizen en tussen twee bussen, de regeringsbussen (SLTB) of de rode private bussen, kiezen. Je kunt met allebei bussen rijden, maar misschien moet je op de private bussen extra voor je bagage betalen en het kan ook echt bezig zijn. Alternatief kan je met een tuktuk of taxi rijden maar er is slim voor de reis met de rijder een prijs af te spreken. Wanneer je een Sri Lankaans rijbewijs hebt of ter plaatse wil maken, kan je ook zelf een tuktuk rijden. Vergelijkbaar, en ook flexibel, kan je ook een privéchauffeur voor de dag huren maar dat is het "duurste" vervoermiddel voor 30 euro per dag. Maar het is ook de flexibelste manier omdat je altijd kunt stoppen.

Ik vind de verschillende mogelijkheden heel interessant maar wanneer ik naar Sri Lanka zou gaan, zou ik met de trein of met een tuktuk rijden. Naar de bus wordt mij altijd misselijk en ik geloof dat de rijders heel snel, en misschien niet zo veilig, rijden om meer mensen, en meer geld, mee te nemen. Wanneer ik met het tuktuk rijd, heb ik ook verse lucht en misschien kan ik het plaatse een rijbewijs maken om zelf te rijden, dat vind ik een leuk idee! Om snel naar andere steden te gaan, zou ik met de trein rijden. Nu ben ik bereid om naar Sri Lanka te gaan!

G Input Text Class Preparation Demo

- Source: VRT NWS (official Belgian news provider)
- Date: 17 September 2026
- URL: https://www.vrt.be/vrtnws/nl/2026/09/17/nieuwe-kattensoort-uantwerpen/

Voor het eerst in 100 jaar nieuwe kattensoort ontdekt: zeldzaam gevlekt dier is kleiner dan een huiskat en werd ontdekt door biologen UAntwerpen

Biologen van de UAntwerpen hebben een nieuwe kattensoort ontdekt. Het gaat om een nieuwe soort tijgerkat, die in de bergbossen van Bolivia leeft. Het is meer dan 100 jaar geleden dat er een nieuwe kattensoort is ontdekt.

Maak kennis met de Leopardus tilcayo, het nieuwste familielid van de tijgerkatten. Het gevlekte katje is iets kleiner dan een huiskat en leeft in de bergbossen van de Andes in Bolivia. De soort blijkt zich 1,4 miljoen jaar geleden te hebben geïsoleerd van de andere tijgerkatten.

Dat blijkt uit genetisch onderzoek van biologen van de UAntwerpen, samen met een team van internationale wetenschappers. Ze ontdekten dat de tilcayo een volwaardige nieuwe kattensoort is.

Verborgen tijgerkat

De vorige keer dat een nieuwe kattensoort werd ontdekt, was in 1923, toen de Uruguayaanse graskat (Leopardus fasciatus) werd beschreven.

"Dat de tilcayo zo lang onopgemerkt is gebleven als nieuwe soort, heeft veel te maken met haar uiterlijk", zegt evolutiebioloog Jonas Lescroart (UAntwerpen), die het onderzoek mee heeft geleid.

"Tijgerkatten zijn kleine, gevlekte katachtigen uit Centraal- en Zuid-Amerika, die op basis van fysieke kenmerken amper onderling te onderscheiden zijn."

"Ook het zeldzame voorkomen en het schuwe karakter van het dier spelen mee in de late ontdekking. De tilcayo leeft in de Boliviaanse Yungas, een grillig, tropisch voorgebergte aan de oostelijke kant van de Andes."

"Dit ecosysteem staat bekend om een zeer grote soortenrijkdom, met fauna en flora die nergens anders voorkomen, maar het gebied is ook moeilijk toegankelijk voor onderzoekers."

Vreemde huiskat

Om te weten hoe de onderzoekers dan toch de nieuwe kattensoort op het spoor zijn gekomen, moeten we 10 jaar terug in de tijd.

"In 2016 werd een gevonden kitten binnengebracht in een Boliviaans lokaal opvangcentrum voor wilde dieren, omdat het toch wel erg vreemd gedrag vertoonde voor een huiskat en er heel anders uitzag", vertelt Lescroart.

Het dier werd toen geïdentificeerd als een noordelijke tijgerkat (Leopardus tigrinus) en kreeg spontaan de koosnaam Tigrino. Het katje trok al snel de aandacht van de Boliviaanse onderzoekers die betrokken waren bij de tijgerkatstudie.

Na een routinebloedafname door de dierenarts werd Tigrino's DNA geanalyseerd. Door die data te vergelijken met stalen van andere kattensoorten, konden de onderzoekers het bewijs leveren dat Tigrino tot een nieuwe kattensoort behoort.

Levend exemplaar

Volgens de onderzoekers is het ook zeer opmerkelijk dat een levend dier is gebruikt om een nieuwe diersoort te beschrijven. Normaal gebeuren de metingen, DNA-analyse en beschrijvingen op een dood dier. Nu gebeurde dat op een levend, verdoofd dier.

"Tigrino zal in het opvangcentrum in Bolivia verblijven tot zijn natuurlijke dood", zegt Lescroart. "Daarna zal het kadaver worden opgenomen in de collectie van het Nationaal Natuurhistorisch Museum van Bolivia."

"De ontdekking van de tilcayo maakt hopelijk meer middelen vrij voor het behoud van populaties tijgerkatten, die zich vaak in een kwetsbare situatie bevinden. Tegelijk vestigen we de aandacht op de ongelooflijke natuurwaarde die Bolivia rijk is."

Uiteindelijk leverden de onderzoekers overtuigend bewijs voor 5 verschillende soorten: de noordelijke en zuidelijke tijgerkat, die al langer als aparte soorten beschouwd werden, de oostelijke tijgerkat en de neveltijgerkat, die recent werden opgewaardeerd van ondersoort tot volwaardige soort, en de tilcayo, die het team vernoemde naar het lokale woord voor tijgerkat.