

1. Background and acknowledgements

NEST is a learner corpus of translations from Norwegian into English, i.e. a supplementary data source for analyses of translation as well as for different aspects of L2 teaching and learning. The corpus project was initiated in 2007, after a long period of teaching translation courses at the University of Oslo and Hedmark University College (by Anne-Line Graedler). The project received financial support from the Arena for Culture and Language (AKS) and the Faculty of Education and Science at Hedmark University College. The initiation, compilation and completion of the NEST corpus project has been inspired and supported by several colleagues from the corpus linguistics research field in Norway: Jarle Ebeling and Stig Johansson, who came up with the idea of creating a translation text corpus in the early 2000s; Sylvi Rørvik, who cooperated with the project design at an early stage; Elizabeth Rasmussen, Hildegunn Dirdal and Sandra Halversen, who all played an essential part in the corpus project by collecting data from their students; Signe Oksefjell Ebeling, who helped by instructing how to align the corpus data using Translation Corpus Aligner (TCA) 2; Sigrid G. Listuen, who worked as an assistant for some of the data preparation; and Knut Hofland, who offered the alignment program (TCA), organized the data and created the user interface.

2. The corpus data

NEST consists of student translation texts submitted during the period 2005 – 2013 by students participating in translation courses as part of regular English studies, not specific translation programs, at tertiary level institutions in Norway.

Material size

The corpus material includes 31 Norwegian source texts (STs) and 349 English translation texts (TTs). Several of the source texts have some slightly different versions, especially because of the size (cf. section 4, the list of all the source texts). Some of the source texts have only one student translation text, but the average is around 10 translation texts per source text.

Most of the source texts are factual prose texts written for a general readership in an everyday language and style. The length of the source texts ranges from 230 – 1,160 words.

In the search interface the source texts occur in parallel with the translation texts, organized sentence by sentence. Hence, the total size of the Norwegian part of the corpus (i.e. all the instances of repeated source texts) is 158,599 words, while the total size of English translation texts in the corpus is 185,826 words.

Data contributors

The NEST data was contributed by 141 students from four different Norwegian universities and colleges: the University of Oslo, the University of Bergen, Sogn og Fjordane University College and Hedmark University College. Most of the students were undergraduates, i.e. in the first or second year at tertiary level English studies. Norwegian university undergraduate students are usually rated as having an 'upper

intermediate’ or ‘advanced’ level proficiency (cf. CEFR). In addition, around 10 per cent of the contributors were in master’s studies. The majority were enrolled at the University of Oslo:

Institutions	1 st year English students (BA level)	2 nd year English students (BA level)	3 rd -4 th year English students (MA level)
University of Oslo	40	58	2
University of Bergen	--	--	14
Hedmark Univ. College	--	24	--
Sogn og Fjordane Univ. College	3	--	--
Total	43	82	16

While the 1st year students’ translations are all based on three source texts, i.e. around 20 translation texts (TT) per source text (ST), the number of translations per source text diminishes at the higher levels, and the average is 11.2 translation texts per source text (some of the source texts have been used both at 2nd and 3rd – 4th year level, cf. section 4. below):

	Source texts (Norwegian)	Translation texts (English)	Tts per ST
1 st year English	3	61	20.33
2 nd year English	22	254	12.09
3 rd – 4 th year English	7	34	4.25
Total	31	349	

More than 80 per cent of the 141 students have Norwegian as their first language (L1). In addition there are some students with other mother tongues, six of which have English:

Students with Norwegian L1:	113
Students with English L1:	6
Students with other L1s:	9
No information:	13

Student background data

In addition to the translation texts, most of the students have contributed some metadata related to their education and language background, anonymized in accordance with statutory data privacy requirements from the Norwegian Social Science Data Services. The metadata is accessible as Excel and HTML documents from the NEST website. The categories that have been included are:

Abbreviations	
General information	
fileid	Code of a specific file
level	Institution and level of study
sem	Semester when the course was taken
sex	Sex
born	Year of birth
nationality	Nationality
years in N	How long they have lived in Norway, if they are not Norwegian
Language background	
L1	First language (mother tongue)
home lang	Language spoken at home, if different from their mother tongue
prim school lang	Main language of instruction during primary school
sec school lang	Main language of instruction during secondary school
foreign lang 1, 2, ...	Other foreign language(s) in addition to English
level fl1, fl2, ...	Proficiency level in other foreign language(s)
Study background	
years of study	Years of university level studies
inst type	Type of institution (university or university college)
degree	Type of education or degree (BA, MA, teacher education, ...)
Background in English	
years of Eng vgs	Years of English study in upper secondary school
written grade Eng vgs	Final grade (written) in English in upper secondary school
oral grade Eng vgs	Final grade (oral) in English in upper secondary school
semesters of Eng study	Number of semesters of English studies at university level
completed ECTS Eng	Completed study points in English studies (ECTS credits)
preferred Eng var	Preferred standard variety of written English
time in Eng environ	Time spent in English-speaking environment with Eng. used daily
details Eng environ	Where and when in English-speaking environment
Background in translation	
semesters of trsl study	Number of semesters of translation studies at university level
completed ECTS trsl	Completed study points in translation studies (ECTS credits)
job in trsl	Have had a job involving translation
trls experience	If yes, the nature of the job, languages, how long, etc.
id	Student ID

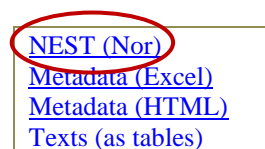
3. Searches

The corpus texts have been aligned and organized by the use of the IMS Open Corpus Workbench (CWB). All of the source texts and translation texts, [Texts \(as tables\)](#), as well as the student metadata,

[Metadata \(Excel\)](#) or [Metadata \(HTML\)](#), are accessible from the web page. Searches can be made of whole or parts of words or word combinations (1-3 words), both in Norwegian and English.

How to search

To perform a simple word search, the relevant section of the corpus , [NEST \(Eng\)](#) or [NEST \(Nor\)](#), must be chosen on the left hand side above the search box.



Type in a single word (e.g. *person*) in the search field and hit the ‘Search’ button. This will produce a result screen with a table of all the sentences where the word *person* occurs in the translation texts (left hand column) and the corresponding sentences in the source texts (right hand column). Up to three single words in three separate search fields can be chosen for the same search.

A screenshot of the search interface. On the left, there is a green sidebar with a 'Search' button circled in red, an 'Ignore case' checkbox checked, and a 'RESET FORM' button. The main search area is yellow and titled 'English' with a UK flag. It contains three search fields. The first field has a dropdown menu set to 'End' and the text 'person' entered, which is circled in red. The second field has a dropdown set to 'Any' and the text 'the'. The third field has a dropdown set to 'Any' and is empty. There are also sections for '2nd word' and '3rd word' with 'Whole word' dropdowns and 'Any' POS dropdowns.

Links to the whole source and translation text in which the sentence occurs are included at the far end of the right hand side of the result table.

The search can be limited to include only a ‘Whole word’, e.g. *person*. The search item can also be chosen as the ‘Start’ of a word (which can extract e.g. *person, personal, personally, persons, etc.*), the ‘End’ of a word, or any ‘Part of a word’ by selecting from the menu to the left of the search field.

A screenshot of the search interface showing the dropdown menu for the first search field open. The menu options are 'Whole word', 'Whole word', 'Start', 'End', and 'Part of'. The 'Whole word' option is highlighted in blue. The search field contains the text 'person'. The rest of the interface is the same as in the previous screenshot.

To have all letters of the search word matching exactly, including case, the check mark (✓) in the box below the search button must be removed.

Specific words can also be filtered out from the search by including them in the field 'Stop words' (e.g. a search for all words starting with *person*, but excluding the forms *personal* and *personally*).

The following categories of student metadata can be used to narrow down and specify the searches:

- Sex
- Year of birth
- Nationality
- How long they have lived in Norway, if they are not Norwegian
- First language (mother tongue), with several abbreviations (cf. https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes):

All = all 1 st languages	NO = Norwegian
BE = Belarusian	PL = Polish
BS = Bosnian	RU = Russian
CHE = Chechen	UK = Ukrainian
DE = German	UR = Urdu
EN = English	VI = Vietnamese
ES = Spanish	XH = Xhosa
Ewe = Niger-Congo	

The screenshot shows the search interface with the following filter options:

- Corpus: NEST (Eng)
- Sex: All
- Born: All
- Nationality: All
- Years in N: All
- L1: All, BE, BS, CHE, DE, EN, ES, Ewe, NO, NO,BS, PL, RU, UK,RU, UR, VI, XH

Below the filters, there are two search boxes: English (with a UK flag) and Norwegian (with a NO flag). The English search box contains the word "person" and the search type is set to "Whole word".

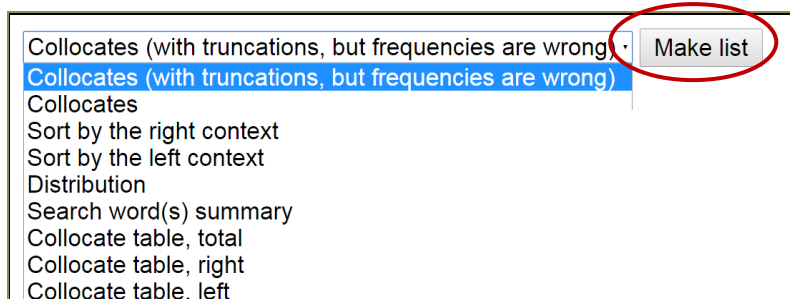
Search results

Besides copying and pasting, the search results can be saved by downloading the data as an Excel document.

The screenshot shows the search results page with the following information:

- Search query: [word="person"%c]:NESTO [word="#"%c];
- Total: 39
- Collocates: [dropdown menu]
- Make list: [button]
- Download to Excel: [button, circled in red]
- Search results:
 - " Among the early pioneers we can mention Moses, Noah (the first **person** clim
 - Petrarch and Henrik Ibsen, the last one yet fundamentally in an abstract sense "
 - Norwegian metaphysician, author and mountaineer. (NEST_Fje_133en.s38)
 - Having completed the educational years at Englebakken high school should be
 - quality on a young **person**. (NEST_EngSP_135en.s7)
 - It is supposed to be a stamp of quality for a young **person** to have completed on

In addition to the direct search results, the search program includes collocation, sorting and distribution of the data. These functions are available from the menu at the top of the search result page. When a function has been chosen, the button next to the menu will organize the data results according to the different choices.



Distribution

The distribution shall give relative frequencies for searches from metadata information, which one can use in the search expression; cf. [Translation Corpus Aligner, version 2](#) (+ [Hofland, K. \(1996\), A Program for Aligning English and Norwegian Sentences](#)).

4. Source texts (ST)

As mentioned above, the corpus material has 31 different Norwegian source texts (ST) which are included in the table below. There is a brief description of the ST's content and genre. In addition, you can see how many target texts (TT) there are for each of the ST, the student level at which the text was used, the full ST titles, and which source texts that have different versions (mostly about the size).

	TT ¹	Level	ST title	ST content and genre
Aul	39	Year 2	<i>Aulestad</i>	Excerpt from a tourist brochure about Aulestad, former home of the Norwegian writer B. Bjørnson, now a museum.
Aul2	14	Year 2	(two ST versions)	
Ber	15	Year 2	<i>Bereiste bunader</i>	Based on a news magazine article about Norwegian historical fashion and costumes (M. Sæbø, <i>Ny Tid</i> , May 11, 2007)
Duo	9	MA	<i>Duo med Scott</i>	Excerpt from the novel <i>Duo med Scott</i> (B. Hagemann, 2012)
Els	3	Year 2	<i>Elsket vi Amerika</i>	Research magazine article about a project on the social relationship between Norway and the US (J.W. Løvhaug, <i>Apollon</i> , 2003, no. 3)
Eng	14	Year 2	<i>Englebakken vgs</i> (five ST versions)	Based on an information leaflet about a Norwegian upper secondary school.
EngSP	3			
EngID	1			
Eng05	3			
Eng05a	1			
Fak	17 2	Year 2 MA	<i>Fakta om Finland</i>	Excerpt from the novel <i>Fakta om Finland</i> (E. Loe, 2001)
Fje	8	Year 2	<i>En fjellvandrers</i>	Excerpt from a magazine article about mountain hiking (T. Wormstrand, <i>Folkevett</i> , 2004, no. 4)

¹ Number of Target Texts in the corpus

bekjennelser				
Fod	1	Year 2	<i>Fødselen</i>	Excerpt from an official information leaflet about pregnancy and birth, from the Norwegian public health service.
Gri	19	Year 2	<i>Grip</i> (two ST versions)	Based on a newspaper article about a Norwegian island fishing village (ØYAVIS – lokalavis for Midsund)
Gri2	15			
Han	3	Year 2	<i>Håndvask og temperaturkontroll</i>	Web text about hygiene and food safety in day-care centers for children in Bergen (Matportalen.no, 2008)
	4	MA		
Har	2	Year 2	<i>Harry våknet</i> (two ST versions)	Excerpt from the novel <i>Kakerlakkene</i> (J. Nesbø, 1998)
Hav	7			
Imo	5	Year 1	<i>I moder jords navn</i>	Based on a magazine article about oil companies in Amazon rain forest (K.-E. Helle, <i>Folkevevt</i> , 2008, no. 2)
Inf	10	MA	<i>Influenza</i>	Text about influenza (The Norwegian Institute of Public Health, 2007)
Inn	2	Year 2	<i>Inngang 3 til paradiset</i>	Three pieces of short 'prose poetry' from <i>Loop</i> (G. Dahle, 2002)
Ins	1	Year 2	<i>Informasjonssamfunnet</i>	Excerpt from a manuscript for a speech given by a Norwegian minister about the development of the information society (J. Lilletun, March 1998)
Kan	7	Year 2	<i>Kan laks lide</i>	Excerpt from a research magazine article about animal welfare and fish farming (A.K.L. Thorolvsen, <i>Forskning</i> , Feb. 2000)
Kli	35	Year 1	<i>Klima i finanskrisens tid</i>	Based on a newspaper article about climate change and global financial crisis (O. Storeng, <i>Aftenposten</i> , Nov. 11, 2008)
Kri	14	Year 2	<i>Krigstidens hverdag</i>	Excerpt from a newspaper article about everyday life in Norway during the Second World War (G. Hjeltnes, <i>Aftenposten</i> , 1990)
Kun	18	Year 1	<i>Kunsten å drikke te</i>	Based on a magazine article about Chinese tea culture (A. Storrønningen, <i>Folkevevt</i> , 2005, no. 6)
Mai	2	Year 2	<i>Maihaugen</i>	Excerpt from a tourist brochure about Maihaugen, a Norwegian outdoor museum with old workshops.
Mil	1	MA	<i>En miljøvennlig jul</i>	Based on a text about environmentally friendly Christmas celebration (C. Leffler, <i>Framtiden i våre hender</i> , 2008)
Nat	15	Year 2	<i>Naturlige valg</i>	Based on a magazine article about healthy food (K.S. Værnes, <i>Folkevevt</i> 2000, no. 2)
Nor	14	Year 2	<i>Nordlys</i>	Excerpt from a documentary book about northern lights, <i>Nordlyset – kulturarv og vitenskap</i> (A. Brekke & A. Egeland, 1994)
Ogn	4	MA	<i>Og Norge</i>	Excerpt from the novel <i>L</i> (E. Loe, 1999)
Opp	13	Year 2	<i>Oppe paa Aulestad</i>	Excerpt from two documentary books about the Norwegian writer B. Bjørnson, <i>I Bjørnstjernes hus</i> , (J.E. Ågotnes, 1992) and <i>Til bords med Bjørnstjerne Bjørnson</i> (J.E. Ågotnes, 1995)
Spe	1	Year 2	<i>Spekk, løgn og hvalfangst</i>	Excerpt from a magazine article about Norwegian whaling (D. Stenerud, <i>Folkevevt</i> , 2003, no. 6)
Sta	3	Year 1	<i>Standardtekst 1</i>	Adapted from a magazine article about snow (T. Wormstrand, <i>Folkevevt</i> , 2003, no. 6)
Str	6	MA	<i>Strøm</i>	Based on a text about saving electricity (Ø. Løtveit, <i>Framtiden i våre hender</i> , 2010)
Sva	6	Year 2	<i>Svartedauden</i>	Excerpt from a documentary book about the Black Death, <i>Svartedauden</i> (Y. Ustvedt, 1985)

Vak	2	Year 2	<i>Vakkert valentinsfrimerke (two ST versions)</i>	Based on a magazine article about postage stamp design (A.C.F. Olsen, <i>ps</i> , 2000, no. 1)
Vak2	1			
Van	9	Year 2	<i>Et vannskille</i>	Excerpt from a newspaper article about the Kursk submarine disaster in Russia (P.N. Waage, <i>Dagbladet</i> , Aug. 22, 2000)
Total	349			

I recommend to do NEST corpus research!

Professor Anne-Line Graedler

anneline.graedler@inn.no

Department of Humanities, Faculty of Education



INLAND NORWAY
UNIVERSITY
OF APPLIED SCIENCES