Apples – Journal of Applied Language Studies
Vol. 8, 3, 2014, 21–38

The Corpus of Advanced Learner Finnish (LAS2): Database and toolkit to study academic learner Finnish

Ilmari Ivaska, University of Turku & University of Washington

This paper introduces the Corpus of Advanced Learner Finnish (LAS2), one of the existing corpora of learner Finnish. The corpus was started at the University of Turku in 2007, and the initial motivation for its collection was to make it possible to deal with novel linguistic challenges posed by academic immigration and to contribute to corpus linguistics, Finnish linguistics and the study of second language acquisition. This paper describes the typological standpoint of the LAS2, its position with respect to other corpora of learner Finnish, the compilation criteria, the annotation applied and the workflow implemented. The corpus consists of three subcorpora of written academic texts of non-native speakers of Finnish. The subcorpora are 1) texts for examination purposes, 2) texts for publishing and graduating purposes, and 3) texts for studying and learning purposes. The informants either study or work in Finnish within academia in Finland. When available, the data has been collected longitudinally. A reference corpus for each subcorpus written by native speakers has also been compiled. Three query tools designed within the framework of the LAS2 are also introduced. These tools enable queries based on any combinations of the linguistic annotation. They can also be used to analyse the typical inner or cotextual variation of any user-specified linguistic node or to create frequency lists of multiword units defined at any level of the annotation. The queries can be limited to a user-specified subset of the data.

Keywords: learner language, Finnish as a second language, corpus typology, annotation, corpus tools

1 Introduction

The ability to communicate, be productive and creative in a language other than one's first learned language has become a necessity in a great deal of professional and academic communication. Despite the ever-growing prevalence of the use of English, there are still a wide range of professions and duties in which thorough mastering of the languages of the society in question is an unconditional necessity. The government of Finland has explicitly underscored the need for greater occupational immigration, and the emphasis is on facilitating the acknowledgement and applicability of existing professional

Corresponding author's email: ilmari.ivaska@utu.fi
ISSN: 1457-9863
Publisher: Centre for Applied Language Studies, University of Jyväskylä

http://apples.jyu.fi

education (Ministry of Labour 2006). Language learning is considered one of the key actions in this process (ibid. 19). Thus, there has been a need for suitable data for systematic analysis of the typicalities and challenges of Finnish as a second language that would also cover the more advanced levels of the learning process and language used in an academic context, and including both quantitative and qualitative components.

A widely recognised position in the study of learner language is to acknowledge the existence of interlanguage in its own terms, not only in comparison to the native variety (Selinker 1972). Nevertheless, contrasting the language used by learners with various language backgrounds as well as with the native variety is a natural way to focus on the features that characterise the language of non-native users (contrastive interlanguage analysis, cf. Granger 1996). This might highlight phenomena potentially common to learning any language (learning universals, e.g. Gentner 2006), phenomena connected to earlier acquired languages (crosslinguistic influences, e.g. Jarvis 2000), or phenomena connected to the language being acquired (target language specific features, e.g. Martin 1995). Again, in order to conduct both quantitative and qualitative systematic analysis to cover advanced learner Finnish used in an academic context, a suitable collection of data was needed.

This paper introduces an ongoing corpus project of written texts of Finnish as a Second language – the Corpus of Advanced Learner Finnish, located at the University of Turku (in Finnish, *Edistyneiden suomenoppijoiden korpus*, henceforth LAS2) (e.g. Ivaska & Siitonen 2009). The aims of the project can be divided into two partial goals. First, a database of academic learner Finnish has been designed, compiled and made available for the purposes of the academic community. The database can serve the needs of linguistic scholars in the field of language acquisition in general, as well as those focusing on Finnish as a second language. The corpus can also be used in the study of language variation. Second, together with producing the actual corpus, the project also aims at contributing to the methodological solutions applied in corpus studies of learner language as well as those of different language varieties in general. This is reflected in, besides the research done in the project, the various query options created to facilitate the research and to broaden the picture a corpus can give about the language use. The tools offer possibilities to approach the data from either a corpus-based or corpus-driven position.

In this paper, I will describe the current corpus (section 2) and locate it typologically among the learner corpora, with particular focus on its relations to other corpora of Finnish (section 3). I will also shed light on the structure and the annotation scheme as well as on the compilation criteria and the workflow. (Section 4). Then, I will introduce the corpus tools available and discuss briefly some potential research designs made possible by the tools (section 5).

2 LAS2 in Figures and Features

Table 1. LAS2 in figures and characterising features (January 2014)

	LAS2
Size	631,402 tokens of raw text in total 640 text units of raw text in total 288,046 tokens annotated (exam questions, tables, example sentences etc. excluded) 476 text units of annotated material
Compilation begun	2007
Different first languages	15
Collection format	Electronic and handwritten
Subcorpora	Three typical academic genres[1]: 1) texts for examination purposes (exam essays etc.); 2) texts for publishing and graduating purposes (thesis and article manuscripts etc.); 3) texts for studying and learning purposes (course papers, learning diaries etc.)
Proficiency level	Higher intermediate or advanced level So far two texts from each informant have been assessed based on the CEFR. The distribution is following: B1: 4%, B2: 45%, C1: 55%, and C2: 6%.
Annotation	Lemmatisation; part-of-speech annotation; morphological annotation; syntactic annotation; possibility for error annotation

The compilation of the LAS2 corpus began in 2007, and it is an ongoing process. The data has been divided into three subcorpora, and it is widely annotated in terms of the language background of the writer, the diachronic location of the text in the timespan of each respective writer's data collection, and in terms of several linguistic features. The first version of the LAS2 consists primarily of texts mostly dealing with the humanities, with a special focus on linguistics. Table 1 describes the contents of the LAS2 in terms of its current size (January 2014) and in terms of various qualitative features. To maximize the comparability to other learner Finnish corpora, the description follows that of the International Corpus of Learner Finnish (ICLFI, Jantunen 2011).

3 Typological Position in the Field of Corpora

Corpora and corpus approaches towards learner language have been emerging for several decades (Granger et al. 2002), but it was not until the latter part of the 2000's that the compilation of the first corpora consisting of learner Finnish took place (for summary, cf. Jantunen & Piltonen 2009). There are currently six corpora that contain at least a component of learner Finnish. Besides the LAS2 corpus, there is the International Corpus of Learner Finnish (ICLFI, Jantunen 2011), the Finnish National Foreign Language Certificate Corpus (YKI[2]), the CEFLING corpus[3], the TOPLING corpus[4] and the Dialuki Corpus[5]. The nature of the data in these different corpora differs in some respects, and they can be seen

to complement each other. While the other corpora consist of texts produced by learners living in Finland, the ICLFI is collected from students of Finnish living outside Finland, and while the LAS2 contains academic texts from advanced learners, the other corpora include texts from various proficiency levels and various genres. The ICLFI, the YKI and the CEFLING allow pseudo-longitudinal studies based on the subcorpora of different proficiency levels whereas the LAS2 and the TOPLING are collected longitudinally. In the LAS2 the time span is usually between one to four years, depending on the informant, while in the TOPLING the span is three years. Additionally, while the other corpora contain texts written by adult learners, the CEFLING and the TOPLING consist of texts written by elementary and high school students.

Jantunen (2011: 88–92) proposes, based on earlier corpus typologies (Atkins et al. 1992; Lehtinen et al. 1995; Laviosa-Braithwaite 1996; Granger 2007), a set of variables by which the nature of the various aspects of a learner corpus can be defined. Defining different corpora using the same variables increases their intercomparability, and I will apply the same set of variables in defining the LAS2 corpus. In contrast with Jantunen, however, I do not describe genre, register and medium dimensions separately. In terms of Halliday and Hasan's definition of register, for example, register consists of the field, mode and tenor of a given text, which in turn are described by the text's situational function and the producer's purposive activity, the production channel and the level of preparation, the genre and the set of relevant social relations participating in the communication (Halliday & Hasan 1976). Similarly, Swales (1990: 58) defines genres as "class[es] of communicative events, the members of which share some set of communicative purposes […] recognized by the expert members of the parent discourse community." According to Swales, "[i]n addition to purpose, exemplars of a genre exhibit various patterns of similarity in terms of structure, style, content and intended audience." Thus, all the three dimensions seem to overlap and intertwine resulting in a tangled web of interacting and correlating features. For this inherent inseparability I describe genre, register and medium dimensions all together.

Genre, register and medium dimensions: LAS2 consists of three genres typical for academic writing (for the definition of genre, see footnote 1), and the corpus is divided into three respective subcorpora: 1) texts for examination purposes (e.g. exam essays); 2) texts for publishing and graduating purposes (e.g. thesis and article manuscripts); 3) texts for studying and learning purposes (e.g. course papers, learning diaries). Texts in the subcorpora 1) and 3) are typically of a private nature, as they are usually intended to be read solely by the instructor. Texts in the subcorpus 2) are written knowing that the final version of the resulting text will be available to a wider audience. Texts in subcorpus 1) have typically been written by hand in invigilated settings and the time for writing the text has been limited. Exams included in this subcorpus have not been taken primarily to assess the language skills, and the primary purpose for writing the texts has been to prove the producer's knowledge of the studied content. The texts in the other two subcorpora have been written in unsupervised settings (e.g. home) without time limitations, typically on a computer, and with access to other sources. The primary purpose for producing the texts in the subcorpus 2) has been to conduct a systematic academic study and to prove the producer's ability to do so. In the subcorpus 3), the primary purposes for the text production have been to reveal explicitly producer's thinking processes and

well-argued opinions related to the studied content and to prove producer's knowledge of the content.

Theme dimension: So far the data has been collected from students of humanities with a focus on linguistics, and the texts in the LAS2 deal thematically mostly with the humanities, particularly linguistics. In the next phase of the project, data collection will be broadened to cover different academic fields and several disciplines, including economics, engineering, medicine, natural sciences, and social sciences.

Language dimension: All the texts in the LAS2 are in Finnish.

Variety dimension[6]: The LAS2 corpus consists of several varieties. On the one hand, there are texts produced by non-native speakers and texts produced by native speakers that can be compared. On the other hand, there are texts from learners with different first languages, which makes it possible to conduct L1-specific studies and comparisons between the different language backgrounds.

Translation dimension: All the texts in the LAS2 are originally written in Finnish, and possible translation exercises have been excluded from the corpus. As Jantunen (2011: 91) points out, however, it is impossible to judge to what extent learners produce the studied language by translating it from other languages they know. Additionally, a great deal of the texts refer to earlier studies that can be written in any language.

Time dimension: The data in the LAS2 has so far been longitudinal in nature, as it has been collected during the studies that last between one to four years.

Sample dimension: In subcorpora one and three, each exam and each paper constitute one text unit of the corpus. In subcorpus two, an article manuscript constitutes one text, whereas each chapter of a thesis constitutes one text unit, as the first drafts are often written during a long period of time and returned to the thesis supervisor separately as the project goes on.

Annotation dimension: LAS2 data is annotated extensively, as the data is lemmatised and annotated grammatically in terms of parts-of-speech, morphological forms and syntactic functions (further details in section 4).

First language dimension: The informants of the LAS2 have various language backgrounds, and as the collection continues, the amount of first languages covered increases all the time. Currently LAS2 contains the language productions of learners with the following first languages: Czech, English, Erzya, Estonian, German, Hungarian, Icelandic, Japanese, Komi, Lithuanian, Polish, Russian, Slovak, Swedish and Udmurt. Furthermore, reference data from native speakers of Finnish has been collected and made available alongside the L2 data.

Proficiency level dimension: Based on the production context of academic discourse, the texts in the LAS2 can usually be said to successfully "pass […] on information or give […] reasons in support of or against a particular point of view" (CEFR: 27). Thus, based on the functional descriptions in the Common European Framework of Reference in language skills, the texts represent usually

the B2 level (vantage level of independent language user), or higher (for level descriptions, ibid. 23–31). So far two texts from each informant have been assessed based on the CEFR and are all divided between B1 and C2: (B1: 4%, B2: 45%, C1: 55%, and C2: 6%).

Learning context dimension: The learning histories of the informants in the LAS2 are heterogeneous. Some of the informants have begun to study Finnish outside Finland, while others have studied Finnish only in Finland. At the time of data collection all the informants study or work within academia in Finland.

Learning method dimension: The learning methods in LAS2 are heterogeneous. On the one hand, all the informants live in Finland at the time of data collection, which enhances natural acquisition, but on the other hand, all the informants have at some point also attended formal language classes either in Finland or elsewhere.

Additionally, there is one further feature to take into account in the LAS2. Following Sinclair's typological definition of corpora, most of the learner language corpora are special corpora, as the texts are often collected in a supervised setting and for the purposes of a corpus (cf. Sinclair 1996). The LAS2, however, differs from most of the learner language corpora, as the texts have been written for genuine academic purposes and not for the purposes of corpus compilation or formal language proficiency assessment. It is likely that advanced learners of a language do pay attention also to the quality of the language they produce and take advantage of their metacognitive skills and other linguistic knowledge they possess even when their language skills are not formally assessed. This is especially likely, as the data has so far been collected mostly from the students of linguistics. The fact that the texts have been produced primarily to fulfil tasks common to everyone within the academia can still be seen to distinguish the data from the texts produced primarily for language learning and/or assessing purposes.

4 Structure, Annotation and Workflow

All the data of the LAS2 corpus will be widely annotated both in terms of textual structure and linguistic features. The data is stored in XML format to ensure the platform independence of the corpus and to maximise the flexibility for later changes, applications and user interfaces. The format follows the TEI guidelines (TEI) applied to Finnish text corpora, such as the Syntax Archives (LaX) and the Morphosyntactic Database of Mikael Agricola's work (Inaba 2007).

The data is stored in text units (<text>), which are labelled in terms of the informant ID (inf), text genre (tt), text unit ID (te), L1 of the informant (l1), the highest (ylin_cefr) and the lowest (alin_cefr) assessment of the informant and, when available, time in months from the first collected text unit of the respective informant (lo). Each text unit is further structured into divisions ( ), paragraphs (<p>), sentences (<s>), clauses (<cl>) and words (<w>). To describe the original text as authentically as possible but at the same time keep the search results reliable, some structural units are labelled separately. These include punctuation, exam questions, linguistic examples, tables and listings, titles, and references. In other words, they are not considered in the query results or word

counts but can be seen when the data is being looked at. Figure 1 depicts the basic annotation structure of the corpus data and the labelling used.

```
Corpus <las2>
    Text unit <text inf="" tt="" te="" lo="" l1="" alin_cefr="" ylin_cefr="">
        Typographic division <div> (exam answer, manuscript or paper subsection)
            Typographic paragraph <p>
                Typographic sentence <s>
                    Grammatical clause <cl type="" com="">
                        Typographic word <w lemma="" pos="" mrp="" fun="" com="">
```

Figure 1. The structure of the LAS2 corpus data and the labelling used

The linguistic annotation is located at the clause and word levels of the annotation structure, and technically it is stored as attributes of the respective nodes. For the clause level, the annotation consists of the modal clause types (type), and for the word level it consists of lemmas (lemma), parts-of-speech (pos), morphological forms (mrp) and syntactic functions (fun). Both levels also enable an optional comment annotation (com), in which the annotator can add a comment on possible errors or potential ambiguities. The LAS2 data is not error-annotated but it could be done in the future in a data-driven manner based on these error comments. To facilitate the comparison between different corpora even further, the aim is to annotate the LAS2 data also in terms of the syntactic dependency structure (Haverinen et al. 2014). Figure 2 gives an example of an annotated sentence of the LAS2 corpus.

```
<s>
<cl type="affdecl" com="">
<w lemma="jotta" pos="cnj" mrp="" fun="lauseyhd" com="">Jotta</w>
<w lemma="voida" pos="v" mrp="fin cond pres sg3" fun="pred" com="">voisi</w>
<w lemma="liittää" pos="v" mrp="inf1" fun="pred2" com="">liittää</w>
<w lemma="sijapääte" pos="n" mrp="sg gen" fun="npobj" com="">sijapäätteen</w>
</cl>
<cl type="affdecl" com="">
<w lemma="olla" pos="v" mrp="fin ind pres sg3" fun="pred" com="">on</w>
<w lemma="muodostaa" pos="v" mrp="pcp1 pass sg nom" fun="pred2" com="">muodostettava</w>
<w lemma="vokaalivartalo" pos="n" mrp="sg nom" fun="npobj" com="">vokaalivartalo</w>
</cl>
<cl type="affdecl" com="">
<w lemma="ja" pos="cnj" mrp="" fun="lauseyhd" com="">ja</w>
<w lemma="pääte" pos="n" mrp="sg gen" fun="nmod" com="">päätteen</w>
<w lemma="liittyminen" pos="n" mrp="sg part" fun="npobj" com="">liittymistä</w>
<w lemma="voida" pos="v" mrp="fin ind pres sg3" fun="pred" com="">voi</w>
<w lemma="seurata" pos="v" mrp="fin ind pres sg3" fun="pred2" com="fin_inf">seuraa</w>
<w lemma="astevaihtelu" pos="n" mrp="sg nom" fun="npsubj" com=""> astevaihtelu</w>
<pn>.</pn>
</cl>
</s>
```

Figure 2. Annotated form of the sentence *Jotta voisi liittää sijapäätteen on muodostettava vokaalivartalo ja päätteen liittymistä voi *seuraa astevaihtelu*." 'In order to add a case ending, a vowel stem needs to be formed, and the addition may cause consonant gradation' from the LAS2 corpus.

Each text unit goes through several processing stages to end up in the structural representation described above. Typical workflow in LAS2 consists of a series of five processes that are in part sequential and in part overlapping. This workflow is described in Figure 3.

Once the informants have given permission to use their texts and they are collected, the texts are converted into plain text files that include structural annotation. Also, the original texts are stored. All the informants also fill in a background questionnaire (cf. Appendix) on their study history, previous and current linguistic circumstances and a self-reflective language proficiency evaluation. The background data is linked to each text unit.

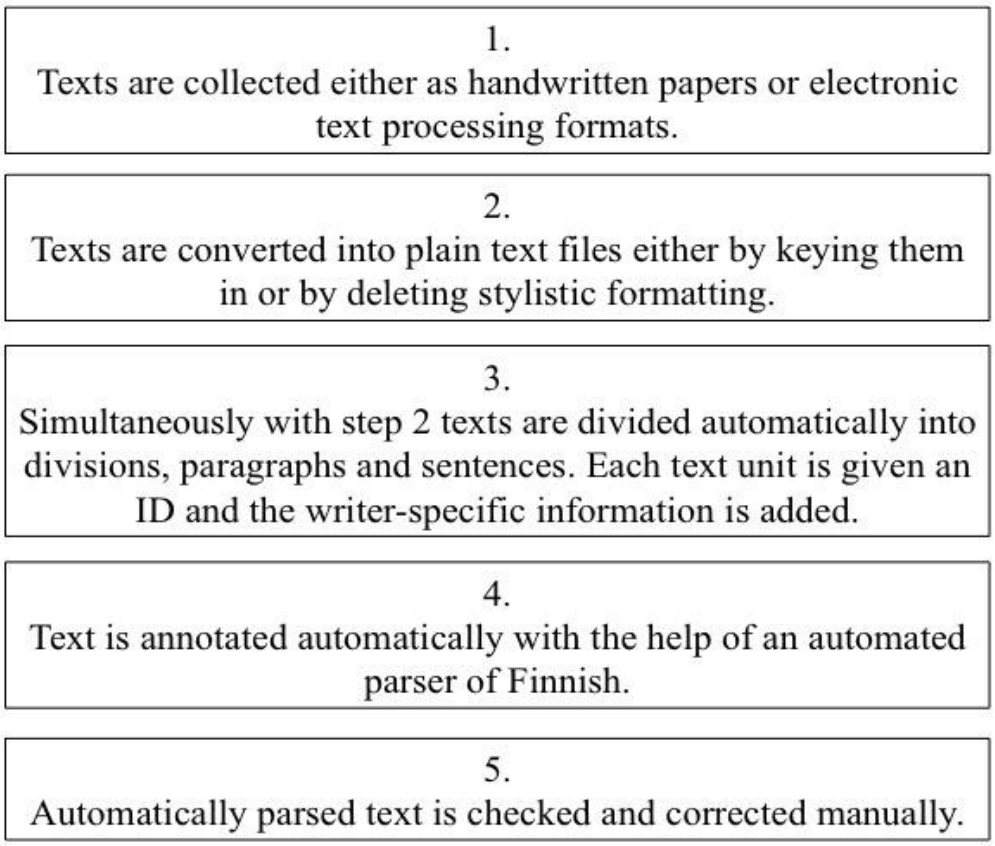

Figure 3. Five steps of the work flow process of a text unit in LAS2

So far, the linguistic annotation has been done semi-automatically, and each occurring word form is annotated manually in terms of its most probable lemma, part-of-speech and morphological form. Then, the information is added to the data and it is annotated syntactically with a probabilistic parser created specifically for the purposes of the LAS2. To ease the labour intensity of the process and to better the results especially in homonymic word forms, the LAS2 team is investigating the use of a parser created by Turku Language Technology Group (e.g. Haverinen et al. 2014), which in turn implements an open-source morphological tool of Finnish called Omorfi[7] and can be used to assign each word its lemma, part-of-speech, morphological form, and the dependency structure of the clause following the widely implemented Stanford Dependency scheme (de Marneffe & Manning 2008). Syntactic functions will also be annotated with the LAS2 parser in the future.

5 Corpus Use

The LAS2 is freely available for the research community, and its use is not dependent on commercial software. As the focus of the project is also to broaden the methodological applications possible in the corpus studies of learner language in general, and Finnish as a second language in particular, software of three different query tools was designed and built for the use of the LAS2: 1) a collostructional query tool, 2) a cotextual query tool and 3) a feature frequency query tool. All the software was written in Java and it can be run in the command line environment[8]. Before the actual query, the user can specify the data to be taken into account in the queries based on the background variables. It can be narrowed down based on the nativeness of the informant, the L1 of the informant, the subcorpus, the specific informant, or any combination of these. As all the data in the LAS2 consists of texts by advanced learners, the informant backgrounds are highly heterogeneous in terms of their study history and the time spent in Finland, among others. Thus, it would not do justice to each individual informant to categorise variables like these into filterable values. Instead, they can be taken into account by orientating oneself with the background description of individual informants.

The collostructional tool allows typical search term based queries of any amount of defined word units. The search terms can be defined at any level of the annotation or their combinations, and also negative search terms are possible. The span of context can be defined to either clause or sentence, but each result can also be backtracked to its original broader textual context. The order of the defined words in the context can be defined as free, relational or consecutive. Thus, possible query definitions could be, for example:

- all the clauses that contain the lemma *saattaa* 'may' and a grammatical subject in any order;
- all the clauses that contain a passive predicate verb and a grammatical object in partitive in this order but not necessarily consecutively;
- all the sentences that contain a noun immediately followed by a verb immediately followed by an adjective.

The search terms are inclusive, and they return all the occurrences that meet the given criteria. For instance, if user queries for all the words in the nominative case, the results contain all the occurrences of words in the nominative, regardless of the other morphological information included in the annotation.

In addition to extracting and returning the occurrences that correspond to the search terms, the collostructional tool also produces and returns a collostructional report of each defined word unit. In other words, it conveys the frequencies of all the different lemmas, parts-of-speech, morphological forms and syntactic functions of each defined word unit, which makes it possible to analyse the typical inner variation (cf. Francis 1993) or collostructional preferences (cf. Stefanowitch & Gries 2003). For example, if the data is queried for all the clauses with grammatical objects, the profile will convey the lexical distribution of all the objects, together with the distributions of different parts-of-speech and different morphological forms acting as objects. The distributional results can also be saved in a platform-independent file format for further processing.

The cotextual tool can be used to extract the typical collocational or colligational features (e.g. Sinclair 1991: 112, 115) of any user-defined node. The node is defined the same way as the collostructional tool. As a result the tool produces a distributional profile of the lemmas, parts-of-speech, morphological forms and syntactic functions preceding the node and following it. Currently, the analysed cotext can be defined as one-, two- or three-word cotext, and the distributions can be looked at separately in each word slot or their combinations. In other words, the possible cotextual profiles produced are:

- the frequency distribution of different lemmas, parts-of-speech, morphological forms and syntactic functions in the preceding and following two-word cotext of the node;
- the frequency distribution of different lemmas, parts-of-speech, morphological forms and syntactic functions one word before and after the node and two words before and after the node.

If there are several words in the node and their order is defined as either free or relational, the cotextual profile will be produced for all the words separately. Again, the distributional results can also be saved in a separate file.

The feature frequency tool can be used to produce frequency lists of the features represented in the corpus. The idea is similar to the word list tools provided in various corpus programs (e.g. Anthony 2012; Scott 2012). In the LAS2, the frequencies can be extracted at any level of the annotation, and it can be used to count multiword units of the length defined by the user. An overall frequency threshold can also be given for an occurrence to be included in the final frequency counts. What is more, the multiword units can be defined either as clusters of consecutive words (e.g. Stubbs & Barth 2003: 62) or as skip-grams, so that the words need not constitute a continuous sequence, as long as they are in the same order and adjacent to each other (Guthrie et al. 2006). Currently, the maximum distance of the words can be defined as being located in the same clause or in the same sentence. For example, Figure 4 depicts a clause consisting of four words *liikkuva*, *väki*, *tarvitsi* and *nimeä*, and the annotation attached with the clause. When two-word units are counted from the clause as continuous word chains, it contains three two-word clusters: *Liikkuva väki*, *väki tarvitsi* and *tarvitsi nimeä*. If the frequencies are counted from skip-grams, however, it consists of six different 2-word units: *Liikkuva väki*, *Liikkuva tarvitsi*, *Liikkuva nimeä*, *väki tarvitsi*, *väki nimeä* and *tarvitsi nimeä*. To improve the statistical applicability of the data, the counts are done separately for each text, and the frequencies are then normalised into occurrences of 1,000 tokens. The counts can be done at any level of the annotation, and the results can be saved in a separate file.

Text:	Liikkuva	väki	tarvitsi	nimeä
Lemma:	liikkua	väki	tarvita	nimi
POS:	V	N	V	N
Mrp:	pcp1 sg nom	sg nom	fin ind pret sg3	sg part
Fun:	nmod	npsubj	pred	npobj
Engl:	'to move'	'people'	'to need'	'name'
	'moving people needed a name'			

Figure 4. Clause *Liikkuva väki tarvitsi nimeä* and the annotation attached to the words

The results of any of the tools can be used in corpus-driven approaches that take advantage of various data-mining techniques, such as statistical keyword analysis (e.g. Scott & Tribble 2006: 58-59), keystructure analysis (Ivaska & Siitonen 2011; Ivaska 2012) and multidimensional analysis (e.g. Biber et al. 2002). Thus, the LAS2 can be used to trace typicalities of any subset of data or to find the common and differing features between different subsets.

To maximise the accessibility of the corpus, the LAS2 group is currently investigating the options for publishing the corpus also online via a user-friendlier graphical interface. A likely option is to publish the corpus in the Language Bank of Finland, which is coordinated by the FIN-CLARIN consortium, a part of the pan-European CLARIN project that aims at easier inter-accessibility and inter-compatibility of the existing language resources[9].

6 Conclusion

This paper introduced the initial motivation for collecting the Corpus of Advanced Learner Finnish and its typological position. It also described the compilation criteria and the annotation of the corpus together with an outline of the workflow implemented. In addition, it gave an overview of the corpus tools offered and examples of possible use. As the LAS2 is an ongoing project, the corpus is in a state of constant change. The size of the data increases all the time, and needs for fine-tuning the details of the process occur every so often. What is more, current and prospective technical implementations always affect the way the corpus can and will be used. Thus, a detailed description and documentation of the criteria are of major importance. Despite the incompleteness of the LAS2 corpus, it should be, and has been, used throughout the project. Thorough piloting and thus applied adjustments play an important role in locating the user needs and the underlying potential of a language resource. Additionally, it can reveal important features of the language forms it covers and, thus, it helps to formulate fruitful research questions for future research.

Besides the accessibility of the corpus, there are also other challenges that require attention also in the future work within the LAS2. The corpus is thematically limited, and while it makes inter-corpora comparisons more accurate, it does at the same time partly limit the generalizability of the results (Jantunen 2011: 101). The intention is to broaden the theme dimension of the corpus to different academic fields and several disciplines. This would allow a

new array of research designs of a comparative nature to study the actual effects of different thematic foci, in addition to the currently possible comparisons of the effects of nativeness, different first languages and different idiolects. What is more, linguistic proficiency level evaluation of all the data in the LAS2 could add another interesting layer both as a background variable and as a possible explaining factor. The current partial evaluation gives a clear overview of the proficiency level of the data as a whole but it does not allow for comparing the different proficiency levels with each other. To conclude, the resources within learner corpus research in general and that of Finnish in particular are very limited. As Jantunen (ibid.) points out, it is in the interest of everyone in the field to maximise the comparability of different corpora by standardising, harmonising and documenting well and publicly all aspects of the compilation criteria, the annotation scheme and the workflow.

Endnotes

1 The concept of genre is problematic and it has been used to refer to various phenomena (cf. Devitt 2004). In this article, the term refers to the functional purpose or the task the text in question is intended to fulfil. It is thus used the same way as the term *genre* in Atkins et al. (1992). In his article written in Finnish, Jantunen (2011) refers to the same concept by terms *genre* and *tekstilaji*.
2 http://yki-korpus.jyu.fi/
3 http://www.jyu.fi/cefling
4 http://www.jyu.fi/topling
5 https://www.jyu.fi/hum/laitokset/solki/tutkimus/projektit/dialuki/en
6 Here, the term 'variety' is used the same way Jantunen (2011) uses it. Thus, it defines a given linguistic form according to a certain background variable, such as the nativeness of the producer or the first language of the producer.
7 http://code.google.com/p/omorfi/
8 The data and the software are available by contacting the LAS2 group.
9 http://www.clarin.eu/external

References

Anthony, L. 2012. *AntConc* (Version 3.3.5). Tokyo: Waseda University.

Atkins, S., J. Clear & N. Ostler 1992. Corpus design criteria. *Literary and Linguistic Computing*, 7 (1), 1–16.

Biber, D., S. Conrad, R. Reppen, P. Byrd & M. Helt 2002. Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, 36 (1), 9–48.

CEFR = *Common European Framework for Languages: Learning, Teaching, Assessment* 2006. Cambridge: Cambridge University Press.

de Marneffe, M. C. & C. Manning 2008. Stanford typed dependencies representation. In *Proceedings of COLING'08 Workshop on Cross-Framed and Cross-Domain Parser Evaluation, Manchester, 23 August, 2008,* 1–8. [Retrieved March 20, 2014]. Available at http://lingo.stanford.edu/events/08/pe/proceedings.pdf

Devitt, A. 2004. *Writing Genres*. Carbondale: Southern Illinois University Press.

Francis, G. 1993. A corpus-driven approach to grammar – principles, methods and examples. In M. Baker, G. Francis, & E. Tognini-Bonelli (eds.), *Text and Technology. In Honour of John Sinclair*. Amsterdam: John Benjamins, 137–156.

Gentner, D. 2006. Why verbs are hard to learn. In K. Hirsh-Pasek & R. Golinkoff (eds.), *Action meets word. How children learn verbs*. Oxford: Oxford University Press, 544–564.

Granger, S. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (eds.), *Languages in Contrast*. Lund: Lund University Press, 37–51.

Granger, S. 2007. A bird's-eye view of learner corpus research. In W. Teubert & R. Krishnamurthy (eds.), *Corpus Linguistics: Critical Concepts in Linguistics*. Vol. 2. London: Routledge, 44–72.

Granger, S., E. Dagneaux & F. Meunier 2002. *International Corpus of Learner English. Handbook and CD-ROM*. Centre for English Corpus Linguistics: Presses Universitaires de Louvain.

Guthrie, D., B. Allison, W. Liu, L. Guthrie & Y. Wilks 2006. A closer look at skip-gram modelling. In *Proceedings of Fifth international Conference on Language Resources and Evaluation (LREC), Genoa, 24–26 May, 2006,* 1222–1225. [Retrieved March 20, 2014]. Available at http://www.lrec-conf.org/proceedings/lrec2006/

Halliday, M.A.K. & R. Hasan 1976. *Cohesion in English*. London: Longman.

Haverinen, K., J. Nyblom, T. Viljanen, V. Laippala, S. Kohonen, A. Missilä, S. Ojala, T. Salakoski & F. Ginter 2014. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48 (3), 493–531. [Retrieved March 20, 2014]. Available at http://dx.doi.org/10.1007/s10579-013-9244-1

Inaba, N. 2007. Mikael Agricolan teokset tietokannan muodossa. (The works of Mikael Agricola in the form of a Database.) In K. Häkkinen & T. Vaittinen (eds.) *Agricolan aika*. (*The Age of Agricola*.) Helsinki: BTJ, 147–161.

Ivaska, I. 2012. Keystructure analysis of formally defined structures of learner Finnish. Paper presented at the *Learner Language, Learner Corpora,* University of Oulu, 5–6 October, 2012.

Ivaska, I. & K. Siitonen 2009. Syntaktisesti koodattu oppijankielen korpus: mahdollisuuksia ja ongelmia. (Syntactically encoded corpus of learner language. Opportunities and challenges.) In P. Eslon, & K. Õim (eds.), *Korpusuuringute metodoloogia ja märgendamise probleemid*. (*The Methodology of Corpus Research and Challenges in Annotation.*) Tallinn: Tallinna Ülikool, 54–71.

Ivaska, I. & K. Siitonen 2011. Avainrakenneanalyysi: Tapa tutkia oppijankielen lauserakennetta korpusvetoisesti. (Key-structure analysis. A way to study clause structure using a corpus-driven approach.) *AFinLA-e* 3, 35–47. [Retrieved January 24, 2014]. Available at http://ojs.tsv.fi/index.php/afinla/issue/view/694.

Jantunen J. H. 2011. Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. (International Corpus of Learner Finnish (ICLFI): typology, variables and annotation.) *Lähivõrdlusi – Lähivertailuja* 21, 86–105.

Jantunen, J. H. & S. Piltonen 2009. Oppijansuomen ja -viron sähköiset tutkimusaineistot. (Electronic research materials of learner Finnish and learner Estonian.) *Virittäjä* 113 (3), 449–458.

Jarvis, S. 2000. Methodological rigor in the study of transfer: identifying L1 influence in the interlanguage lexicon. *Language Learning*, 50 (2), 245–309.

Laviosa-Braithwaite, S. 1996. *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation*. Manchester: UMIST.

LaX = The Syntax Archives. School of Languages and Translation Studies of the University of Turku. Turku. [Retrieved March 22, 2013]. Available at http://syntaxarchives.suo.utu.fi.

Lehtinen, M., P. Karvonen & T. Rahikainen 1995. *Tekstikorpukset. Raportti tekstikorpusten koostamisperiaatteista ja nykysuomen tekstiaineistojen tarpeellisuudesta Kotimaisten kielten tutkimuskeskuksessa*. (*Text corpora. A report regarding the compilation principles and the necessity of the text data of modern Finnish at the Research Institute for the Languages of Finland*.) Helsinki: Kotimaisten kielten tutkimuskeskus.

Martin, M. 1995. *The Map and The Rope. Finnish Nominal Inflection as a Learning Target*. Jyväskylä: University of Jyväskylä.

Ministry of Labour 2006. *Hallituksen maahanmuuttopoliittinen ohjelma*. (*Government programme regarding immigration policies*.) Työhallinnon julkaisuja 371. Helsinki: Työministeriö.

Scott, M. 2012. *WordSmith Tools* (version 6). Liverpool: Lexical Analysis Software.

Scott, M. & C. Tribble 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

Selinker, L. 1972. Interlanguage. *International Review of Applied Linguistics*, 10, 209–231.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. 1996. *EAGLES. Preliminary Recommendations on Corpus Typology*. Cambridge Approaches to Linguistics. [Retrieved November 17, 2011]. Available at http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html

Stefanowitsch, A. & S. Gries 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8 (2), 209–243.

Stubbs, M. & I. Barth 2003. Using recurrent phrases as text-type discriminators: A quantitative method and some findings. *Functions of Language*, 10 (1), 61–104.

Swales, J. 1990. *Genre Analysis. English in Academic and Research Settings.* Cambridge: Cambridge University Press.

TEI = TEI Consortium, (eds.) *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. [P5, version 2.3.0]. TEI Consortium. [Retrieved February 6, 2013]. Available at http://www.tei-c.org/Guidelines/P5/

YKI = Finnish National Foreign Language Certificate Corpus [online database]. University of Jyväskylä. Centre for Applied Language Studies [administrator and producer]. [Retrieved February 5, 2012]. Available at http://yki-korpus.jyu.fi

Appendix: Taustatietolomake (Background information form)

Tiedot on kerätty _____, ja ne kuvaavat informantin tilannetta tuona ajankohtana.
päivämäärä

(*The information has been collected* _____ *and it describes the situation at that point in time.*)
date

Tietoja käytetään nimettöminä Edistyneet suomenoppijat -materiaalin (LAS2) yhteydessä.

(*The information will be used anonymously as a part of the Corpus of Advanced Learner Finnish (LAS2).*)

SYNTYMÄVUOSI (*YEAR OF BIRTH*): _____

SUKUPUOLI (*GENDER*): nainen (*female*) ( ), mies (*male*) ( )

SYNTYMÄMAA JA -PAIKKA (*COUNRTY AND PLACE OF BIRTH*):
_____

ENSIKIELI (*FIRST LANGUAGE*):
_____

OLEN KAKSIKIELINEN – kielet ovat _____ ja _____.
(*I AM BILINGUAL* – *the languages are* _____ *and* _____.)

EN OLE KAKSIKIELINEN, mutta toinen vanhemmistani/isovanhemmistani on suomalainen ja olen aina puhunut vähän myös suomea
(*I AM NOT BILINGUAL* *but one of my parents/grandparents is Finnish and I have always spoken also some Finnish*)
_____
(Kerro, kuka/ketkä sukulaisistasi ovat suomalaisia)
(*Tell who of your relatives is/are Finnish*)

EN OLE KAKSIKIELINEN (***I AM NOT BILINGUAL***) ( )

<u>SUOMEN KIELEN OPISKELEMINEN JA OPETTAMINEN</u>
<u>(FINNISH STUDIES AND TEACHING EXPERIENCE)</u>

Kuinka kauan olet asunut Suomessa?
6 kk ( ), 7 – 12 kk ( ), 13 – 18 kk ( ), 19 – 24 kk ( ), 2 – 3 v ( ), 3 – 4 v ( ), 5 – 6 v ( ), 7 – 8 v ( ), yli 8, ___ vuotta
(*How long have you been living in Finland?*
6 months ( ), *7 – 12 months* ( ), *13 – 18 months* ( ), *19 – 24 months* ( ), *2 – 3 years* ( ), *3 – 4 years* ( ), *5 – 6 years* ( ), *7 – 8 years* ( ), *over 8, ___ years*)

Kuinka kauan olet opiskellut suomea Suomessa?
(*How long have you been studying Finnish?*)

_____

Mitä olet suorittanut? (Kielikursseja, Suomen kielen ja kulttuurin perusopinnot, Suomen kielen ja kulttuurin aineopinnot, muuta, mitä?)
(*What have you studied?* *Language courses, lower-division courses in Finnish language and Culture, upper-division courses in Finnish language and culture, what else?*) _____
_____

Oletko opiskellut suomea jossain muualla? Missä ja kuinka kauan?
(*Have you studied Finnish somewhere else? Where and for how long?*)
En ole opiskellut suomea muualla. (*I have not studied Finnish elsewhere*) ( )
Olen opiskellut suomea muualla __ vuotta. (*I have studied Finnish elsewhere __ years.*)
Missä? (*Where?*)

_____
_____

Oletko opettanut suomea? Missä ja kuinka kauan?
(*Have you taught Finnish? Where and for how long?*)
En ole opettanut suomea. (*I have not taught Finnish.*) ( )
Olen opettanut suomen kieltä __ vuotta. (*I have taught Finnish __ years.*)
Missä? (*Where?*)

_____
Kuinka kauan? (*For how long?*) _____
Mitä kieltä puhuville? (*What was the first language of the students?*) _____

SUOMEN KIELEN KÄYTTÖ JA SUOMENKIELISET KONTAKTIT
(*USE OF FINNISH AND FINNISH CONTACTS*)

Mitä kieliä kotonasi käytetään? (*What languages are used at your home?*)

_____

Mitä kieliä itse puhut kotona? (*What languages do you speak at home?*)

_____

Mitä kieliä puhut kodin ulkopuolella säännöllisesti?
(*What languages do you speak regularly outside your home?*)

_____

Suomenkieliset kontaktini ovat seuraavanlaisia:
(*My Finnish contacts are following*:)
kasvokkainen keskustelu (*face-to-face conversation*) ( )
puhelinkeskustelu (*telephone conversation*) ( )
tekstiviesti (*text message*) ( )
sähköposti (*email*) ( )
kirje (*letter*) ( )
Joku muu, mikä? (*Something else, what?*)

_____

Henkilöt, joiden kanssa puhun suomea (*People with whom I speak Finnish are*)
Puoliso (*spouse*) ( )
joku muu perheenjäsen (*other family member*) ( )
Kuka? (*Who?*) _____
omat sukulaiset (*my own relatives*) ( )
puolison sukulaiset (*relatives of my spouse*) ( )
naapurit (*neighbours*) ( )
työtoverit (*colleagues*) ( )
henkilöt harrastusten parista (*contacts from hobbies*) ( )
muut (*other*) ( )

Minulla on tuttavia, joiden kanssa puhun vain suomea. ( )
(*I know people with whom I speak only Finnish*.)
1–5 ( ) , 6–10 ( ) , 11–15 ( ) , 16– _____ ( )

SUOMEN KIELEN LUKEMINEN JA KIRJOITTAMINEN
(*READING AND WRITING FINNISH*)

Olen alkanut kirjoittaa suomeksi säännöllisesti vuonna _____.
(*I have begun to regularly write in Finnish in year _____*.)

Kirjoitan suomeksi myös muuta kuin opiskeluun liittyvää. ( )
(*I also write other than study-related texts in Finnish*.)
Mitä? (*What?*)
_____

Olen elämässäni lukenut suomeksi (*In my life I have read in Finnish*)

kaunokirjallisuutta n. _____ sivua (*approximately ____ pages of fiction*)
tieteellistä kirjallisuutta n. _____ sivua (*approximately ____ pages of academic literature*)

Luen suomenkielistä sanomalehteä (*I read newspapers in Finnish*)
joka päivä (*every day*) ( )
kerran viikossa (*once a week*) ( )
harvemmin (*less frequently*) ( )

Olen kääntänyt suomesta äidinkieleeni n. _____ sivua tekstiä.
(*I have translated from Finnish into my first language about _____ pages of text.*)

Olen kääntänyt eri kielistä suomeen n. _____ sivua tekstiä.
(*I have translated from other languages into Finnish about _____ pages of text.*)

ARVIO OMASTA SUOMEN KIELEN TAIDOSTA
(*SELF-EVALUATION OF YOUR OWN FINNISH SKILLS*)

Numeroi seuraavat kielitaidon osa-alueet sen mukaan, millä alueella arvioit oman suomen kielen taitosi parhaaksi (1.), toisiksi parhaaksi (2.) jne. Merkitse samalla taitotasolla olevat samalla järjestysnumerolla.
(*Estimate the following aspects of your language skills in a ranked order from the strongest (1.), the second strongest (2,) and so forth. The aspects you estimate to be equally strong should be marked with the same number.*)

puhuminen (*speaking*) ( )
puheen ymmärtäminen (*listening*) ( )
kirjoittaminen (*writing*) ( )
tekstin ymmärtäminen (*reading*) ( )
rakenteiden hallinta (*grammar*) ( )
sanaston hallinta (*vocabulary*) ( )

YLEINEN KIELITAITO
(*OTHER LANGUAGE SKILLS*)

Mitä kieliä puhut? (*What languages do you speak?*)
Puhun seuraavia kieliä (*I speak following languages*):

_____

Mitä kieliä ymmärrät? (*What languages do you understand?*)
Ymmärrän seuraavia kieliä (*I understand following languages*):

_____

MUUTA HUOMAUTETTAVAA:
(*OTHER THINGS YOU WOULD LIKE TO MENTION:*)

_____
_____
_____

Received December 9, 2013
Revision received September 8, 2014
Accepted November 25, 2014