CEFR-nivåer och svenska flerordsuttryck

Lindström Tiedemann, Therese

Svensk-österbottniska samfundet
2022

Lindström Tiedemann , T , Alfter , D & Volodina , E 2022 , CEFR-nivåer och svenska flerordsuttryck . i S Björklund , B Haagensen , M Nordman & A Westerlund (red) , Svenskan i Finland 19 : Föredrag vid den nittonde sammankomsten för beskrivningen av svenskan i þÿ F i n l a n d , V a s a d e n 6 7 m a j 2 0 2 1 . S k r i f t e r u t g i v n a a v S v e n s k - Ö s t e r b o t t 82 , Svenskan i Finland , Svensk-österbottniska samfundet , Vasa , s. 218-233 , Svenskan i Finland , 06/05/2021 .

http://hdl.handle.net/10138/346472

unspecified
publishedVersion

CEFR-nivåer och svenska flerordsuttryck

Therese Lindström Tiedemann, David Alfter & Elena Volodina

1 INLEDNING[1]

När vi lär oss ett nytt språk ska vi inte bara lära oss enstaka ord och hur vi använder dessa, utan vi måste också lära oss vilka ordkombinationer som är ”fasta uttryck” till betydelsen (t.ex. *hälsa på någon*) eller till formen (t.ex. *lättare sagt än gjort*) eller båda delarna (t.ex. *huller om buller*). Enligt en del studier kan dessa uttryck utgöra så mycket som 50 % av vokabulären i ett språk som förstaspråk (L1) eller ännu mer (Jackendoff 1997; Erman 2007, 28). Men det är möjligt att de är vanligare i vardagligt språk och talspråk (Prentice & Sköldberg 2013). Flerordsenheter kan vara problematiska för andraspråkstalare (Nesselhauf 2003, 223) till och med på avancerad nivå (jfr Pawley & Syder 1983; Wray & Perkins 2000; Nesselhauf 2003; Prentice 2010). Samtidigt är de en helt nödvändig del av språket (Nesselhauf 2003, 223) och kan utmärka andraspråkstalarna som icke-modersmålstalare (Pawley & Syder 1983; Wray 2002). Flerordsuttryck är alltså en värdefull del av andraspråkskompetensen (se även Paquot 2019) och något som är viktigt att studera hur vi på bästa sätt introducerar för L2-talaren och om de kan kopplas till nivåer i bedömning.

I den här studien presenterar vi resultat kring förståelsen av flerordsuttryck i svenska som andraspråk i relation till färdighetsnivåerna enligt *Gemensam Europeisk Referensram för Språk* (GERS eller CEFR, *Common European Framework of Reference*) (COE 2001; 2018; Skolverket 2009; Utbildningsstyrelsen 2018). Vår analys består av många olika typer av data:

1. Data från ett crowdsourcingexperiment som vi gjort med andraspråkstalare (B1-nivå och uppåt, nedan kallade *L2-talare*) och lärare, forskare och bedömare (nedan kallade *L2-lärare*) i svenska som andraspråk (L2).
2. Förekomsterna i kursbokskorpusen *Coctaill* (*Corpus of CEFR-based Textbooks as Input for Learner Levels' modelling*, Volodina m.fl. 2014).
3. Förekomsterna i andraspråkstalares produktion i *SweLL-pilot*-korpusen (*Swedish Learner Language,* Volodina m.fl. 2016a).
4. Explicit nivåannotering av tre L2-lärare med god kännedom om CEFR (nedan kallade *CEFR-experter*).
5. Förekomsterna i referenskorpusar för L1-svenska (se vidare i avsnitt 2).

[1] Studien har finansierats av Riksbankens jubileumsfond, och är en del av projektet *Utveckling av lexikala och grammatiska kompetenser i invandrarsvenska* (dnr P17-0716:1). Vi vill också tacka för kommentarer från konferensdeltagare och granskare.

Vi har i tidigare studier kvantitativt jämfört hur L2-talare och L2-lärare ser på flerordsuttryckens receptiva svårighetsgrad på basis av samma crowdsourcing-experiment och samma direkta nivåannotering från CEFR-experter som vi analyserar här. Vi konstaterade då att L2-talare och L2-lärare är relativt överens i sina *relativa* bedömningar enligt kvantitativa analyser (Alfter m.fl. 2020; Alfter m.fl. 2021). Här gör vi en mer kvalitativ analys av delar av datan och undersöker dessutom hur resultaten relaterar till uttryckens förekomster i bruket baserat på olika korpusar.

Den här typen av kvalitativ analys kräver att vi begränsar oss till endast en viss del av alla data från crowdsourcingexperimentet eftersom det blir för komplext att kvalitativt jämföra rankningarna av alla 60 uttrycken i alla tre grupperna mellan de tre deltagargrupperna (60 x 3 x 3 = 540 relativa rankningar plus lika många explicita nivåannoteringar). En kvantitativ analys har som sagt redan gjorts av alla rankningar, men en mer kvalitativ analys behövs för att bättre förstå vad datan säger oss om vilka uttryck som upplevs som lätta eller svåra att förstå och hur pass tillförlitliga de olika metoderna är för att uppskatta färdighetsnivån. Som urval fokuserar vi på de uttryck som rankades som lättast eller svårast av de olika deltagargrupperna. Det kan tyckas som de som borde vara enklast att bestämma och därför som de minst intressanta, men det finns skillnader även här och därför anser vi att det är bäst att först analysera dessa för att bättre förstå hur vi kan nå en så bra nivåannotering som möjligt.

Vårt syfte är att i längden uppnå en metod för nivåannotering av svensk lexis enligt färdighetsnivåer såsom CEFR-skalans för att detta ska kunna användas för inlärning, undervisning och examination. Vi vill förstå vilka uttryck som upplevs som lätta och svåra att förstå och när olika grupper är mest överens om detta. Hur samspelar det med förekomster i kursböcker och med input från allmäntillgängligt skriftligt inflöde, såsom tidnings- och bloggkorpusar? Våra forskningsfrågor är:

1. Hur eniga är de tre deltagargrupperna om vilka uttryck som rankats som lättast och svårast *att förstå* i de olika grupperna av flerordsuttryck?
2. Hur relaterar de lättaste respektive de svåraste uttrycken i rankningsexperimentet till förekomsterna i kursböckerna i COCTAILL?
3. Hur relaterar rankningen av uttrycken till deras förekomst i L2-talares produktion baserat på SweLL-pilotkorpusen?
4. Hur överensstämmer rankningen med CEFR-experternas explicita nivåannotering?
5. Finns det en korrelation med uttryckens förekomst i tidningskorpusar och bloggkorpusar och sålunda med sannolikheten att man kommer i kontakt med dem i bruket utanför språkundervisningen?

Avsnitt 2 beskriver studiens metod och material. I avsnitt 3 presenteras sedan en analys av våra resultat. Avsnitt 4 sammanfattar och diskuterar kort våra resultat och ger förslag på kommande forskning.

2 METOD OCH MATERIAL

Den här studien baserar sig på många olika data som vi försöker relatera till varandra. Vi utgår först och främst från resultat från ett crowdsourcing-experiment[2] kring rankning av hur lätt eller svårt det är för en andraspråkstalare av svenska att *förstå* vissa flerordsuttryck. Det är således *receptiv kunskap* vi fokuserar på.

Crowdsourcingexperimentets design baserades på Čibej m.fl. (u.a.) och flerordsuttrycken valdes ut från Coctaillkorpusen. Coctaillkorpusen är en korpus över sverigesvenska kursböcker i svenska som andraspråk, för vuxna L2-talare. Böckerna har annoterats med CEFR-nivåer (A1–C1) baserat på lärares beskrivning av vilken nivå de används för i undervisning av svenska som andraspråk (Volodina m.fl. 2014). Baserat på Coctaill och inlärarkorpusen SweLL-pilot har vi skapat en ordlista, *Sen*Lex* (Alfter m.fl. u.a.), som är en vidareutveckling av SvaLex och SweLLex (François m.fl. 2016; Volodina m.fl. 2016b). Sen*Lex består av ord och flerordsenheter som förekommer i kursböcker (Coctaill-korpusen) och texter skrivna av L2-talare av svenska (SweLL-pilotkorpusen). Varje ord eller flerordsenhet har med hjälp av Sparv (Borin m.fl. 2016) annoterats med lemma, ordklass, lemgram, betydelse och flerordsenheter. Flerords-enheterna har identifierats på basis av flerordsenheter som finns med i Språkbankens lexikon, *Saldo*, (Borin m.fl. 2013; jfr Borin 2021; jfr Sag m.fl. 2002). Två annoterare[3] har sedan manuellt klassificerat alla flerordsenheter enligt olika typer. Baserat på den klassifikationen har vi sedan valt ut flerordsenheter till de tre olika grupperna i crowdsourcingexperimentet:

> **Grupp 1:** Fasta uttryck, idiom och interjektioner (*lycka till, på pin kiv*)
> **Grupp 2:** Verbala uttryck inkl. reflexiva verb och partikelverb (*tycka om, ta illa upp*)
> **Grupp 3:** Adverbiella, adjektiviska och icke-lexikala uttryck (*i alla fall, på håret*).

Vi valde 60 uttryck från varje grupp, 12 per CEFR-nivå baserat på första förekomsten i kursbokskorpusen Coctaill och vi försökte välja så jämnt som möjligt mellan de olika kategorierna i den manuella annoteringen.

CEFR- nivåerna är främst kopplade till kommunikativa färdigheter och nivåerna bör snarast ses som ett kontinuum (COE 2018, 34, jfr även i relation till färdighet generellt Ortega 2012). Kommunikativa färdigheter innefattar specifika lingvistiska kompetenser såsom vokabulär (2001, 108), vokabulärens omfång och lexikal kontroll (2001, 112). Trots detta kan man säga att vad man kan göra kommunikativt med språket kan göras med olika lexis och grammatik och att det är svårt att koppla CEFR-nivåer till exakta ord och uttryck. Men redan i den ursprungliga CEFR-publikationen (COE 2001, 30) menade författarna att

[2] För närmare beskrivning av experimentet se Alfter m.fl. (2021).
[3] Annoterarna var modersmålstalare av svenska (finlandssvenska). En med en MA i nordiska språk och en var disputerad i nordiska språk.

detaljerade listor över bl.a. vokabulär kunde specificeras per språk (t.ex. *Threshold Level* 1990). De säger explicit att användare av CEFR kan vilja specificera ”vilka lexikala enheter (fasta uttryck och enskilda ord) som inläraren behöver/har förutsättningar att/måste känna igen och/eller använda” (Skolverket 2009, 109; jfr COE 2001, 112) och hur de ska väljas ut och ordnas. CEFR-författarna uppmuntrar till analys som försöker koppla kommunikativa uppgifter i CEFR till specifik vokabulär (2001, 33) och flerordsenheter (*fixed expressions*) och lexikal kompentens behandlas i CEFR (2001, 110–111), inte bara generell kommunikativ färdighet.

Uttryck som baserat på förekomsten i kursböckerna i Coctaill kunde associeras med A1 valdes som uttryck för A1-nivån med målet att undersöka om L2-talare och L2-lärare (lärare, forskare, bedömare i L2-svenska) också såg dessa som bland de enklaste och var eniga om sin rankning. Detta gjordes med förhoppningen om att hitta nya tids- och kostnadseffektiva rankningssätt som även kunde ses som tillförlitliga. För uttryckets kursboksnivå utgick vi från den första förekomsten eftersom det visat sig att relativ förekomst, dispersion och *significant onset of use* inte nödvändigtvis ger mer korrekta nivåer och dessa metoder är mer komplicerade (jfr Alfter 2021, kap. 5–6, Alfter m.fl. 2016 för produktiv nivå; Alfter & Volodina 2018; jfr också Gala m.fl. 2014). Det är förstås möjligt att ord och uttryck förekommer i kursböcker på en nivå där man inte behöver kunna dem produktivt, men kursboksförfattarna antas utgå från att de flesta orden och uttrycken kan läras in, *förstås* och/eller behövas på den nivån. Vi återkommer i analysen och diskussionen till hur pass första förekomst utan beaktan av dispersion kan ge bra nog resultat.

Även tidigare ordlistor har oftast utgått från *receptiv* kompetens och frekvenser i texter skrivna av och för L1-talare. I likhet med *English profile*-projektet, som arbetat med att utveckla *Reference Level Descriptions* (RLD)[4] för engelska, har vi baserat oss på empiriska korpusdata och anser att vi på så sätt kopplar till CEFR:s *can-do*-koncept då uttryck faktiskt förekommer i kursböcker som lärare säger (se Volodina m.fl. 2014) används för undervisning på vissa CEFR-nivåer och som i vissa fall explicit är ämnade för vissa nivåer enligt förlaget. English profile använder dock primärt inlärardata dvs. L2-talares produktion, men de tar också ordlistor i kursböcker i beaktan (jfr Capel 2010; 2012; 2015; Hawkins & Buttery 2010; Hawkins & Filipović 2012). Vi har sett ett intresse i att kunna jämföra receptiva (Coctaill) och produktiva (SweLL-pilot) data. Men liksom English Profile utgår vi från att baserat på informationen om förekomster i inlärardata kan ord och uttryck kopplas till CEFR-nivåer (Capel 2012, 1). Ett annat internationellt projekt som använt kursboksdata för att studera CEFR-nivåer för lexis är CEFRLex-projektet (t.ex. François m.fl. 2016). För svenska har även KELLY-listan länkat ord med CEFR-nivåer, men baserat på frekvenser och dispersion i en L1-webbkorpus och med en indelning över nivåerna enligt lika stora delar (Volodina & Johansson Kokkinakis 2012a; 2012b; Kilgariff m.fl. 2014).

[4] https://www.coe.int/en/web/common-european-framework-reference-languages/cefr-reference-level-descriptions-language-by-language-components-and-forerunners

I den här studien jämför vi nivån i kursböckerna med en explicit rankning av CEFR-experter (jfr Capel 2015, 14 om *Cambridge English Lexicon* som kompilerats av Hindmarsch (1980) främst utifrån intuition). I kalkylark med alla uttrycken ombads de tre CEFR-experterna explicit annotera uttrycken med en CEFR-nivå (A1–C2 eller högre) *efter* att de rankat alla uttrycken i pyBossa (326 x 3 projekt). Varje grupp med 60 uttryck utgjorde grunden för ett av tre crowdsourcingprojekt som vi utförde i Språkbankens installation av pyBossa.

Deltagare i crowdsourcingexperimentet fick miniuppgifter med fyra flerordsuttryck där de skulle bedöma vilket som var lättast *att förstå för en andraspråkstalare av svenska* och vilket som var svårast att förstå. All rankning och bedömning skedde utan kontext, vilket är en onaturlig situation men eftersom kontexten lätt gör att meningen som helhet bedöms använde vi isolerade uttryck. Samma design hade dessutom valts tidigare i Čibej m.fl (u.a.) som vi tog inspiration ifrån. Varje uttryck förseddes med en definition baserad på t.ex. *Svensk ordbok* som kunde tas fram genom att klicka på uttrycket.

I vår analys här fokuserar vi på de sju lättaste och de sju svåraste uttrycken i varje grupp av uttryck och från varje deltagargrupp, vilket ändå innebär fler än 7 x 3 uttryck eftersom grupperna inte är helt eniga. Antalet sju är valt för att inte bli alltför omfattande men ändå omfatta en någorlunda stor andel av vardera gruppen av flerordsuttryck. Det innebär att vi granskar ca 25 % av alla rankningar (14 av totalt 60 per grupp). Rankningen från experimentet jämförs kvalitativt mellan de tre olika deltagargrupperna, men också med förekomster i kursbokskorpusen Coctaill, inlärarkorpusen SweLL-pilot samt med direkta nivåannoteringar av tre CEFR-experter. Till sist jämför vi med bruket i tidningskorpusar och bloggkorpusar, som ett sätt att studera en annan del av inflödet som en L2-talare kan tänkas få. Tidnings- och bloggkorpusarna vi har använt är både finlandssvenska (Hufvudstadsbladet (Hbl) 1991–2014 och Bloggtexter 2006–2013) och sverigesvenska (Göteborgsposten (GP) 1994–2013 inkl. 2 dagar och Bloggmix 1998–2017, samt okänt år). Alla korpusarna finns tillgängliga via Språkbanken Text och sökningarna har gjorts med hjälp av användargränssnittet Korp (Borin m.fl. 2012).

3 RESULTAT OCH ANALYS[5]

Vi presenterar här vår jämförelse av vardera deltagargrupps rankning i crowdsourcingexperimentet. Vi behandlar först de sju lättaste uttrycken (3.1) och sedan de sju svåraste (3.2). I båda avsnitten presenterar vi först jämförelsen mellan deltagargrupperna. Sedan undersöker vi vilka kursboksnivåer som orden förekommer på, och huruvida orden förekommer även i inlärartexter (3.1.1 resp.

[5] En förteckning över de lättaste och svåraste uttrycken, samt hur CEFR-experterna annoterat dem med CEFR-nivåer och deras förekomster i olika korpusar finns på: https://researchportal.helsinki.fi/sv/publications/cefr-niv%C3%A5er-och-svenska-flerordsuttryck

3.2.1). Ibland är det en förvånande skillnad mellan rankningen och förekomstnivån i kursböckerna där den uppskattade nivån kan vara lägre än kursboksnivån. För att bättre förstå detta jämför vi med hur CEFR-experterna bedömde uttryckens nivå i det direkta annoteringsexperimentet där de tilldelade varje uttryck en specifik CEFR-nivå från A1 till C2+ (3.1.2, 3.2.2). I sista avsnittet i varje del jämför vi sedan alla dessa resultat med referenskorpusar för svenska som förstaspråk (3.1.3, 3.2.3).

3.1 De sju lättaste uttrycken

Bland de sju lättaste uttrycken förekom nästan bara uttryck som vi valt ut från A1-nivån i kursböckerna. Men det fanns två uttryck som inte förekom på A1-nivå: *inga problem* (B1) och *logga in* (A2). De olika deltagargrupperna var mest eniga vad gäller verbuttryck (grupp 2) och adverbiella /adjektiviska uttryck (grupp 3) där fem av sju uttryck fanns bland de lättaste sju i alla tre deltagargrupper. Interjektioner m.m. (grupp 1) hade bara tre av sju uttryck gemensamt bland alla tre deltagargrupper. Dessutom var det två uttryck i varje uttrycksgrupp som bara fanns bland de sju lättaste uttrycken i en av deltagargrupperna: *ingen fara, pommes frites*[6]*; hälsa på, ta med; i alla fall, mitt i*.

Med tanke på att Spearmankoefficienten var högst för grupp 1 enligt våra tidigare studier (Alfter m.fl. 2020; 2021) så är det här lite förvånande. Men vi kan notera att det är endast i grupp 1 som vi har uttryck som rankats *exakt* lika av alla tre deltagargrupperna bland de lättaste: *god morgon, god natt* som sågs som de två allra lättaste i alla deltagargrupperna.

3.1.1 Förekomst i kursböcker och andraspråksproduktion

Bland de uttryck som hade rankats bland de lättaste sju av *alla* deltagargrupperna var de flesta från A1-nivå i kursböckerna (11 st., 85 %). Det fanns två undantag *inga problem* (B1) och *logga in* (A2). *Inga problem* förekommer i kursboksdialog på B1-nivå och finns totalt 4 gånger på den nivån, men det förekommer också på högre nivåer (B2, C1). *Logga in* förekommer på A2-nivå, men då bara i en översättningsövning och sedan i några texter på C1-nivå.

Flera (10 st., 77 %) av de lättaste uttrycken återfinns även i inläraruppsatser i SweLL-pilotkorpusen, dvs. används även *produktivt*. De uttryck som alla var eniga om att de var bland de lättaste sju finns dessutom på de lägsta nivåerna i inlärarkorpusen: A1-nivå (4) och A2-nivå (6), men vissa finns inte med i inlärarkorpusen överhuvudtaget (3). Här måste vi dock ha i åtanke att det finns en risk att något missats i inlärartexterna p.g.a. felstavning (jfr t.ex. Stemle m.fl. 2019). Men lemmatiseringen har visat sig vara bra även i inläraruppsatser och inte så

[6] Lånade uttryck undveks men det här togs med eftersom det inte fanns fler uttryck på nivån och det trots allt är ett uttryck som används frekvent i svenska.

annorlunda från kursbokstexterna eller normaliserade versioner av inlärar-uppsatser (Volodina m.fl. 2021).

Att inte alla de lättaste uttrycken finns på A1- eller A2-nivå i inlärartexterna kan ha flera orsaker. Även om dessa uttryck har rankats bland de som är lättast att förstå så kan de vara svåra att lära sig producera. Förekomsterna i inlärartexter påverkas dessutom av de ämnen som inlärarna fått skriva om (se t.ex. Caines & Buttery 2017; Capel 2015, 15). Därför är det mycket intressant att de två uttrycken som inte förekom på A1-nivå i kursböckerna förekommer redan på A2-nivå i inlärarnas egna texter. Så trots att de kommer upp lite senare än A1 i kursböckerna behärskas de relativt tidigt produktivt och kanske då faktiskt förstås ännu tidigare. Här misstänker vi att det kan påverka att det här är uttryck som är vanliga i vardagsspråket (jfr Prentice och Sköldberg 2013 om att flerordsuttryck ofta är vanliga i talspråk och vardagsspråk). Dessutom påverkas *logga in* troligen av att det är ett internationellt uttryck som kan kännas igen från andra språk L2-talaren kan.[7]

3.1.2 Jämförelse med experternas nivåannotering

I direktannoteringen av nivå, som tre CEFR-experter gjorde, rankades 7 (54 %) av de lättaste uttrycken som A1 av *alla* tre CEFR-experterna. Fyra till rankades som A1 av två men A2 av den tredje (*laga mat, tycka om, till höger, inga problem*). Två rankades som A2 av två och A1 av den tredje (*för sent* och *logga in*). Alla var alltså eniga om att *alla* 13 av dessa uttryck var från A-nivå.

Av de två undantagen från tidigare, *logga in* och *inga problem*, var *logga in* ett av de uttryck som rankades som lite svårare (A2) av två CEFR-experter vilket är helt i överensstämmelse med den första kursboksförekomsten. Dessutom var *inga problem* ett av de fyra som rankats som A2 av en av CEFR-experterna och som först förekom på B1-nivå i kursböckerna. Dessa kom alltså lite senare än A1 i kursböckerna och var inte så vanliga där heller, dessutom var de lite svårare enligt den explicita nivåbedömningen av CEFR-experterna, men ändå inte speciellt svåra vilket stämmer bra med den relativa rankningen i crowdsourcing-experimentet. *Inga problem* rankades som bland de svåraste av de lättaste sju i den relativa rankningen (plats 6–7), men *logga in* såg L2-lärarna och CEFR-experterna som bland de allra lättaste (plats 3) medan L2-talarna rankade det som lite svårare (plats 6). Lite förvånande är det att *logga in* hamnade på plats 3 relativt men explicit bedömdes som A2 av två experter. Det här är dock antagligen ett tecken på att den direkta annoteringen är mycket mer subjektiv och mer kognitivt belastande än relativ rankning (för mer diskussion se Alfter m.fl. 2021). Med tanke på att uttrycken upplevdes som så pass lätta i experimentet även om de ibland kom på högre nivåer i kursböcker, undersöker vi om det kan vara relaterat till inflöde från andra kanaler än kursböcker och om vi där kan se något som stöder Prentice och Sköldberg (2013) och deras idé om att flerordsuttryck kan vara vanligare i mer vardagligt språk.

[7] Jfr eng. *log in,* ty. *einloggen*.

3.1.3 Jämförelse med L1-referenskorpusar, tidningstexter och bloggtexter

Merparten av de lättaste uttrycken (10 st., 77 %) används mer i kursbokskorpusen än i tidningstexter. Det här skulle kunna bero på genreskillnaden där tidningstext är en viss genre och kursböcker innehåller flera olika genrer, men kanske allt på de teman som uttrycken kan kopplas ihop med. Kursböckerna utgår specifikt från de teman som kopplas till olika färdighetsnivåer: t.ex. A1 berätta om sig själv, A2 berätta om omgivningen, osv. Tidigare studier har påpekat att hög frekvens i allmänspråkliga korpusar utelämnar vissa viktiga uttryck för inlärare (se t.ex. Volodina & Johansson Kokkinakis 2012b, 10; Capel 2015, 13; François m.fl. 2016). Men uttrycken som förekommer på B1–C1-nivå i kursböckerna (dvs. *logga in, inga problem*) är lite mindre vanliga i kursböcker och infrekventa även i tidningskorpusarna, speciellt *logga in*. Men frekvensen för *logga in* liknar den för *god natt* som alla var eniga om var lätt. Båda är ovanliga i tidningstexter (*god natt* GP 0,7, Hbl 2,8 och *logga in* GP 2,7, Hbl 2).[8] Eftersom tidigare forskning (Prentice & Sköldberg 2013) menat att flerordsuttryck ibland är vanligare i talspråk och vardagligt språk kan det vara intressant också att jämföra med förekomsten i mer informella texter. Bloggkorpusar innehåller åtminstone delvis mer informellt språk än tidningstexter. Det är tydligt att många uttryck verkligen förekommer betydligt mer där än i tidningskorpusar och även mer än i kursböckerna. *Logga in* Bloggmix 15,5, Bloggtexter 8 och *god natt* Bloggmix 26,6, Bloggtexter 8,8.

Varför är det då främst A1-ord som uppfattas som lättast att förstå för en andraspråkstalare? Uttryck på A1-nivå kan förväntas vara vardagsuttryck. De är också oftast frekventa i kursböcker, tidningskorpusar och/eller bloggtexter. Vissa av uttrycken kan dock förväntas vara vanliga främst i tal vilket kan förklara varför många av uttrycken verkar vanligare i kursböckerna än i tidningstexter som kan vara fel genre för den här typen av uttryck, speciellt vad gäller interjektioner, men de kan ändå övas i dialoger i kursböckerna.

3.2 De sju svåraste flerordsenheterna

De olika deltagargrupperna i experimentet var allra mest eniga om de svåraste uttrycken i grupp 1 *interjektioner m.m.*, vilket stämmer väl överens med våra tidigare kvantitativa resultat som visat högst enighet för grupp 1 mätt med Spearmankoefficient (Alfter m.fl. 2021). Fyra av uttrycken var med bland alla gruppernas sju svåraste i grupp 1. I de andra två grupperna var det flera uttryck som fanns med i två av tre deltagargrupper, men endast tre uttryck i grupp 2 och två uttryck i grupp 3 var sådana att *alla* deltagargrupperna var eniga om att det hörde till de svåraste sju.

Deltagargrupperna var helt eniga om att *åka snålskjuts* var det allra svåraste uttrycket i grupp 2 och nästan helt eniga om det svåraste i grupp 1 *på pin kiv* (det blev svårast eller nästsvårast). Intressant nog så fanns det trots den någorlunda

[8] Frekvenser anges i relativa frekvenser per 1 miljon ord.

höga samstämmigheten i grupp 1 två–tre uttryck från varje deltagargrupp som inte kom med i någon av de andra gruppernas svåraste sju.

3.2.1 Förekomst i kursböcker och andraspråksproduktion

De uttryck som rankades bland de svåraste sju förekom i kursböckerna på många olika nivåer och gav således en spretigare bild än de lättaste. Bland de fyra som alla var eniga om i grupp 1 kom *ett* från C1 i kursböckerna (*på pin kiv*), medan resterande tre förekom för första gången på B2-nivå (*veta hut, så det stod härliga till, med nöd och näppe*). I de två andra grupperna fanns det också ett uttryck i vardera från C1 (*åka snålskjuts, på sin höjd*), men de andra uttrycken som alla tre grupper hade placerat bland de sju svåraste kom från B2 (*rabbla upp*) och A2 (*slå läger*) i verbgruppen och från B1 i gruppen med adverbiella och adjektiviska uttryck (*på håret*). Varför blev vissa uttryck från lägre nivåer än C1 i kursböckerna, ändå placerade bland de svåraste sju av alla tre deltagargrupperna? För att försöka förstå det här bättre tar vi en närmare titt på kursboksförekomsterna och undersöker också förekomsterna i andraspråksproduktion genom SweLL-pilotkorpusen.

Inga av de svåraste uttrycken finns med i inlärarkorpusen (SweLL-pilot). Produktiv kompetens ses ofta som något som kommer efter receptiv. Men som påpekades ovan så kan förekomsten i inlärarkorpusen också bero på de ämnen som inlärarna fått skriva uppsatser om (jfr Caines & Buttery 2017). Det skulle dock vara intressant att testa produktionen på annat sätt än genom fri produktion, exempelvis genom vokabulärtest där L2-talare med olika färdighetsnivå ska definiera eller använda uttrycken, men också test där de får definitioner och ska identifiera uttryck t.ex. i flervalsfrågor.

Om vi tittar på de uttryck som grupperna sett som svårast och som ändå förekommer på en klart lägre nivå än C1 i kursbokskorpusen: *slå läger* (A2) och *på håret* (B1), så är de ovanliga över lag i kursbokskorpusen och förekommer inte heller ofta på högre nivåer. *På håret* förekommer tre gånger totalt men endast två gånger i den här betydelsen (B1, C1). Intressant nog finns det med i olika böcker men böcker som enligt lärare används på väldigt olika nivåer (*Svenska utifrån B1, Skrivtrappan C1*[9]). Förekomsterna är slumpmässiga och att de är på olika nivåer gör det svårt att uppskatta den receptiva nivån för uttrycken baserat på förekomsterna i kursbokskorpusen något som borde tas i beaktan i eventuella automatiska rankningar på basis av kursböcker (jfr tidigare experiment som beaktat dispersion, t.ex. Alfter m.fl. 2016). Att de förekommer i flera olika böcker visar på att det är ett uttryck som man kan förväntas lära sig kring den nivån. Det vore intressant att titta närmare på kontexten som flerordsuttrycken förekommer i i de olika läromedlen för att få en bättre förståelse för hur det kan vara att det varierar så vilken nivå på läromedel som de förekommer i:

[9] Fetstil är tillagd i alla exempel om inte annat påpekas. *På håret* finns även med i *Nya mål 3 B2* men i en bokstavlig användning: ”Min världsmästare sitter och suger **på håret,** tänker han…” (berättelse)

(1) det var **på håret** (uppgift, Skrivtrappan, C1)
(2) Många gånger var jag **på håret** att falla, jag viftade med stavarna… (anekdot, Svenska utifrån, B1)

Bruket visar här att exempel (1) snarast är något som L2-talare förväntas kunna använda produktivt på C1-nivå, medan förekomsten i boken på B1-nivå är rent receptiv. Nivåskillnaden stämmer alltså överens med förväntan om att receptiv kompetens ska komma innan produktiv. Med detta i åtanke är det dock intressant att det här ändå setts som ett av de svåraste uttrycken.

Slå läger förekommer likaså sällan i kursböckerna, totalt två gånger och i olika böcker, på olika men angränsande nivåer (*Svenska utifrån A2, På svenska! 2 B1):*

(3) … när man har tältat eller **slagit läger** någonstans (fakta, Svenska utifrån, A2)
(4) De tre männen bestämde sig för att vandra med sina slädar över is och snö [...] till Vitön där de **slog läger** i oktober. (faktaberättelse, På svenska! 2, B1)

Här är båda förekomsterna rent receptiva, och båda från liknande genre. Att de råkar förekomma i kursböcker på olika nivåer är troligen en effekt av att färdighet ska ses som ett kontinuum. Men det är också tydligt att det är svårt att säga något om svårigheten baserat på första förekomsten i kursböcker *om* uttrycket förekommer väldigt sällan och i få böcker samt på olika nivåer i de olika böckerna. När nivåerna är angränsande är det ändå ett tydligt tecken att uttrycket förväntas kunna förstås *kring* en viss nivå, och troligen inte bör ses som mycket lättare eller svårare än så.

Crowdsourcingexperimentet är en relativ rankning, så det är inte sagt att de svåraste ska ses som C1-nivå. Det kan också vara att alla uttrycken i själva verket kan ses som så lätta att förstå att en L2-talare på A1-nivå skulle kunna förstå dem, men att man ändå uppskattar vissa som svårare än andra (jfr tanken om färdighet som ett kontinuum t.ex. Ortega 2012). CEFR är en skala även om nivåerna ofta ses som kategoriska av olika skäl (COE 2018, 34). Ibland förekommer det också att man delar in nivåerna i undernivåer såsom A1.1 och A1.2 (jfr finska nivåskalan Utbildningsstyrelsen 2019; 2021; Hildén & Takala, 2007). Om något ska ses som olika svårt *inom* en CEFR-nivå eller som *olika* CEFR-nivåer kan en relativ rankning inte säga oss. Men vi kan jämföra med den rankning som CEFR-experterna gjorde där de för varje uttryck bedömde på vilken CEFR-nivå uttrycket kunde förstås genom att välja nivå (A1–C2+) i en dropdownmeny för varje uttryck i ett kalkylark.

3.2.2 Jämförelse med experternas nivåannotering

Bland de uttryck som alla grupperna tyckte var bland de sju svåraste varierade den direkta nivåannoteringen från CEFR-experterna mellan A2 och C2+. Vi ser att det är endast ett uttryck som alla tre bedömer som samma nivå, nämligen *på pin kiv* som bedöms som C2+. Intressant med tanke på att den sammantagna poängen från CEFR-experternas crowdsourcingresultat var att *på pin kiv*

hamnade på en andra plats, medan det kom på första plats, dvs. som allra svårast, för L2-talare och L2-lärare. Alla är alltså mycket överens om att det här är ett svårt uttryck. De uttryck som i CEFR-experternas crowdsourcingresultat blev svårast var *så det stod härliga till, åka snålskjuts* och *av och till* vilka i den direkta nivåannoteringen sågs som C1/C2+, B1/C1 respektive C1/C2+. Vi har tidigare visat kvantitativt att CEFR-experterna inte var överens med sig själva i en jämförelse av crowdsourcingrankningen och direktannoteringen (Alfter m.fl. 2021), och här ser vi alltså att detsamma är sant om vi tittar närmare på resultatet för vissa uttryck. Detta beror med all sannolikhet på att uppgifterna är så pass olika och att direktannotering är betydligt mer kognitivt krävande (se vidare diskussion i Alfter m.fl. 2021).

I tre andra fall inom de sju svåraste uttrycken som alla hade med, är alla CEFR-experterna överens om C-nivå: *veta hut, på håret* och *så det stod härliga till*. Mest oeniga är de vad gäller *slå läger* (A2–C1) vilket också hade en markant låg första förekomst i kursböckerna (A2).

3.2.3 Jämförelse med L1-referenskorpusar

Uttryck som har rankats som svåra är ofta ganska ovanliga i kursbokskorpusen och inget av dem förekommer i inlärarkorpusen. De flesta uttrycken är dessutom ganska ovanliga i referenskorpusarna medan vissa av uttrycken förekommer väldigt mycket mer i kursböckerna i jämförelse med referenskorpusarna. Vi ser dock ingen markant skillnad i frekvensen i referenskorpusarna i relation till kursboksnivån för första förekomsten för de olika uttryck som rankats bland de svåraste.

4 SAMMANFATTNING OCH DISKUSSION

Som vi har visat i tidigare studier (Alfter m.fl. 2020; 2021) så är de relativa rankningarna av flerordsuttrycks receptiva nivåer väldigt lika enligt L2-talares och L2-lärares intuitioner om svenska som andraspråk. I den här studien har vi visat att moderat samstämmighet också finns bland de svåraste och lättaste sju i vardera gruppen av flerordsuttryck. Totalt 28 uttryck finns bland dem som av någon av deltagargrupperna har rankats som de lättaste sju uttrycken, vilket ska jämföras med 21 (dvs. 7 x 3 grupper) om grupperna varit helt eniga om uttrycken. Tretton (46,4 %) av alla de lättaste uttrycken fanns i alla tre grupperna. Fem (17,9 %) fanns både i L2-talargruppen och en av de andra grupperna som båda representerar lärare, forskare och bedömare. Det innebär att 18 (64,3 %) uttryck hade minst partiell samstämmighet.

De svåraste uttrycken var deltagargrupperna mindre eniga om, vilket syns i att de svåraste sju ger totalt 35 olika flerordsuttryck (i stället för 7 x 3 = 21 uttryck om de varit helt eniga). Av dessa var dessutom endast nio (25,7 %) desamma i alla tre deltagargrupperna. Och för endast sju (20 %) fanns en enighet mellan L2-talare och en av de andra grupperna, vilket innebär att endast 16 (45,7 %) hade

minst partiell samstämmighet. Deltagargrupperna var alltså klart enigare om de lättaste uttrycken (46,4 % – 20 % total enighet; 64,3 % – 45,7% partiell).

I en jämförelse av rankningen och kursboksförekomsten förekom 18 (18/21, 85,7 %) av L2-talarnas lättaste sju på A1-nivå, 16 (76,2 %) av L2-lärarnas och 15 (71,4 %) av CEFR-experternas. L2-talarnas rankning verkar således stämma bäst överens med kursböckernas för de *lättaste* uttrycken. För de svåraste uttrycken var dock samstämmigheten med kursböckerna svårare att bedöma och det var inte bara uttryck med första förekomst på C1-nivå i kursböckerna som rankades bland de svåraste sju. L2-talarna hade 9 (43 %) uttryck från C1-nivå bland sina svåraste sju, L2-lärarna 11 (52 %) och CEFR-experterna 7 (33 %). L2-lärarna var alltså mest överens med C1 i kursböckerna, men vidare studier skulle behöva säkerställa hur svåra uttryck på högre nivåer kan uppskattas vara, vilket sannolikt relaterar till en större variation i inflödet på högre nivåer. Det här är ett ganska väntat resultat eftersom det är tydligt att det som behandlas på A1-nivå är ett mycket mer begränsat tematiskt område, vilket också torde göra uttrycken mer lika mellan olika böcker och olika klassrum än på C1-nivå. På C1-nivå ska man kunna täcka en mängd olika ämnen och det är ytterst sannolikt att vokabulären varierar mer mellan olika kursböcker och inlärarsituationer och det kan också göra att det finns en risk att L2-talare på den nivån har större variation i sin lexikala kompetens om man ser den i relation till specifika uttryck (jfr COE 2001, 112) som noterar ett fokus på områden av intresse på B2-nivå, men på C-nivå förväntas man lösa produktiva vokabulärproblem genom omskrivningar. Upp till B2-nivå ser man ett fokus på allmän vokabulär (COE 2001, 28)). Det vore intressant att studera vidare hur rankningen skiljer sig mellan kärnuttryck som förekommer oftare och i fler kursböcker på en nivå eller angränsande nivåer, och perifera uttryck som förekommer sällan och endast i enstaka böcker.

Vissa uttryck som förekom sent i kursböckerna rankades ändå som lätta vilket vi misstänker har att göra med att de är talspråkliga, vardagliga uttryck som man stöter på i klassrummet och samhället vilket finner stöd i Prentice och Sköldbergs (2013) konstaterande om att flerordsuttryck ofta är vanliga i dessa genrer. Vi kan se att en del av de lätta uttrycken också förekommer mer i blogg-korpusar än i tidningskorpusar och kursböcker vilket skulle kunna vara ett tecken på att de hör till vardagsspråket. Vissa av de ”lätta” uttrycken är också sådana att deras rankning säkerligen påverkas av tidigare kända språk och hur internationella orden är, som t.ex. *logga in* som förekom först på C1-nivå i text i kursböckerna men i en övning på A2-nivå och det var bland de lättaste sju i alla grupperna och nivåannoterades med A1- och A2-nivå av CEFR-experterna i den direkta nivåannoteringen.

L2-talare och L2-lärare verkar ha en mer enhetlig bild av svårighetsnivån vad gäller *lätta* flerordsuttryck, vilket stämmer väl med vad man kan förvänta sig baserat på vad som är känt om ordkunskapen på olika nivåer (jfr COE 2001). Det skulle vara intressant att studera nivårankning vidare i relation till enkla ord och hela texter (jfr CLAP-projektet[10]). Eftersom vi ser en större oenighet bland de svårare uttrycken både mellan deltagargrupperna och i hur pass uttrycken

10 https://perso.uclouvain.be/magali.paquot/research-projects/

förekommer på C1-nivå i kursböckerna vore det intressant att titta mer på vilken vokabulär som övas i kursböcker och i undervisning på mer avancerade nivåer och hur detta relaterar till bedömning enligt CEFR-nivåer. För de lägsta nivåerna är potentialen att kunna uppskatta nivån baserat på första kursboksförekomsten bra men dispersionen tycks bra att ha i åtanke. L2-talare ger dessutom en bra uppskattning, men även L2-lärare.

De lättaste uttrycken var tydligt mer frekventa i bruket än de svåraste. Många lätta uttryck var speciellt vanliga i bloggar. Med andra ord finns god anledning att vidare undersöka hur bruket i inflödet korrelerar med L2-talares färdighet. Vår studie visar att det här kan göras väl genom en kombination av crowdsourcing och korpusstudier, men lexikala test som testar förståelse i kontext och förmåga att kombinera uttrycken med definitioner, eller definiera dem vore intressanta för jämförelser och validering av nivåer.

LITTERATUR

Alfter, D. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*. PhD Thesis. Data Linguistica 31. Göteborg: Göteborgs universitet.

Alfter, D., Bizzoni, Y., Agebjörn, A., Volodina, E. & Pilán, I. 2016. From Distribution to Labels: A Lexical Proficiency Analysis using Learner Corpora. I: *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition.* SLTC, Umeå. Linköping: Linköping University Electronic Press. S. 1–7.

Alfter, D., Lindström Tiedemann, T. & Volodina, E. 2020. *Expert judgments versus crowdsourcing in ordering multi-word expressions*. Eighth Swedish Language Technology Conference (SLTC). https://gubox.app.box.com/v/SLTC-2020-paper-16.

Alfter, D., Lindström Tiedemann, T. & Volodina, E. 2021. Crowdsourcing Relative Rankings of Multi-Word Expressions: Experts versus Non-Experts. I: *Northern European Journal of Language Technology (NEJLT*), vol. 1 2021. DOI: 10.3384/nejlt.2000-1533.2021.3128.

Alfter, D., Lindström Tiedemann, T. & Volodina, E. u.a. Sen*Lex.

Alfter, D. & Volodina, E. 2018. Towards single word lexical complexity prediction. I: *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. New Orleans, USA. S. 79–88.

Borin, L., Forsberg, M. & Roxendal, J. 2012. Korp – the corpus infrastructure of Språkbanken. I: *Proceedings of LREC 2012*. Istanbul: ELRA. S. 474–478.

Borin, L., Forsberg, M. & Lönngren, L. 2013. SALDO: a touch of yin to WordNet's yang. I: *Language resources and evaluation*, volume 47, issue 4. S. 1191–1211.

Borin, L., Forsberg, M., Hammarstedt, M., Rosen, D., Schäfer, R. & Schumacher, A. 2016. Sparv: Sprakbanken's corpus annotation pipeline infrastructure. I: *The Sixth Swedish Language Technology Conference* (SLTC), Umeå University.

Borin, L. 2021. Multiword expressions – a tough typological nut for Swedish FrameNet++. I: Dannélls, D., Borin, L. & Friberg Heppin, K. (red.), *The Swedish FrameNet++ harmonization, integration, method development and practical language technology applications*. John Benjamins. S. 221–259.

Caines, A. & Buttery, P. 2017. The effect of task and topic on opportunity of use in learner corpora. I: *Learner corpus research: New perspectives and applications*. S. 5–27.

Capel, A. 2010. A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. I: *English Profile Journal* 1 (1). DOI:10.1017/S2041536210000048.

Capel, A. 2012. Completing the *English Vocabulary Profile:* C1 and C2 vocabulary. I: *English Profile Journal* 3 (1). DOI:10.1017/S2041536212000013.

Capel, A. 2015. The English Vocabulary Profile. I: Harrison, J. & Barker, F. (red.), *English Profile Studies 5. English Profile in Practice*. Cambridge: Cambridge University Press. S. 9–27.

CEFRLex u.å. https://cental.uclouvain.be/cefrlex/.

Čibej, J., Alfter, D., Kosem, I. & Volodina, E. u. a. Multi-word expressions and language learning: validity of a crowdsourcing approach.

Council of Europe [COE]. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Council of Europe [COE]. 2018. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors.* Hämtad 19.10.2021. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989.

Erman, B. 2007. Cognitive processes as evidence of the idiom principle. I: *International journal of corpus linguistics*. 12(1).

François, T., Volodina, E., Pilán, I. & Tack, A. 2016. *SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners*. Proceedings of LREC 2016, Slovenia.

Gala, N., François, T., Bernhard, D. & Fairon, C. 2014. Un modèle pour prédire la complexité lexicale et graduer les mots. I: *TALN 2014*. S. 91–102.

Hawkins, J. A. & Buttery, P. 2010. Criterial Features in Learner Corpora: Theory and Illustrations. I: *English Profile Journal*. 1 (1). DOI:10.1017/S2041536210000103.

Hawkins, J. A. & Filipović, L. 2012. Criterial Features in L2 English. I: *English Profile Studies 1*. Cambridge: Cambridge University Press.

Hilden, R. & Takala, S. 2007. Relating descriptors of the Finnish school scale to the CEF overall scales for communicative activities. I: Koskensalo, A., Smeds, J., Kaikkonen, P. & Kohonen, V. (red.), *Foreign languages and multicultural perspectives in the European context = Fremdsprachen und multikulturelle Perspektiven im europäischen Kontext.* Dichtung - Wahrheit - Sprache; Vuosikerta Band 9–10. Lit Verlag. S. 291–300.

Hindmarsch, R. 1980. *Cambridge English Lexicon.* Cambridge: Cambridge University Press.

Jackendoff, R. 1997.*The architecture of the language faculty*. MIT Press.

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Bondi Johannessen, J., Khalil, S., Johansson Kokkinakis, S., Lew, R., Sharoff, S., Vadlapudi, R. & Volodina, E. 2014. Corpus-based vocabulary lists for language learners for nine languages. I: *Language Resources and Evaluation,* 48: 121–163. DOI: 10.1007/s10579-013-9251-2.

Korp u.å. https://spraakbanken.gu.se/korp/.

Nesselhauf, N. 2003. The use of collocations by advanced learners of English and some implications for teaching. I: *Applied linguistics. 24*. S. 223–242.

Ortega, L. 2012. Interlanguage complexity. I: Kortmann, B. & Szmrecsanyi, B. (red.), *Linguistic complexity: Second language acquisition, indigenization, contact.* De Gruyter. S. 127–155.

Paquot, M. 2019. The phraseological dimension in interlanguage complexity research. I: *Second Language Research*, *35*(1). S. 121–145.

Pawley, A. & Syder, F. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. I: Richards, J.C. & Schmidt, R.W. (red.), *Language and Communication.* London: Routledge. S. 203–239.

Prentice, J. 2010. *Käppen i hjulen. Behärskning av svenska konventionaliserade uttryck bland gymnasieelever med varierande språklig bakgrund.* Rapporter i svenska som andraspråk (ROSA 12). Göteborg: Institutet för svenska som andraspråk, Göteborgs universitet.

Prentice, J. & Sköldberg, E. 2013. Flerordsenheter – ur ett andraspråksperspektiv. I: Hyltenstam, K. & Lindberg, I. (red.), *Svenska som andraspråk – i forskning, undervisning och samhälle.* Lund: Studentlitteratur. S. 197–220.

PyBossa u.å. https://pybossa.com/.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. 2002. Multiword expressions: A pain in the neck for NLP. I: *International Conference on Intelligent Text Processing and Computational Linguistics*. S. 1–15. Springer, Berlin, Heidelberg.

Skolverket 2009. *Gemensam europeisk referensram för språk: lärande, undervisning och bedömning*. Skolverket. https://www.skolverket.se/publikationer?id=2144.

Stemle, E. W., Boyd, A., Jansen, M., Lindström Tiedemann, T., Mikelić Preradović, N., Rosen, A., Rosén, D. & Volodina, E. 2019. Working together towards an ideal infrastructure for language learner corpora. I: Abel, A., Glaznieks, A., Lyding, V. & Nicolas, L. (red.), *Widening the Scope of Learner Corpus Research: Selected Papers from the Fourth Learner Corpus Research Conference*. Corpora and Language in Use, Presses universitaires de Louvain, Louvain-la-Neuve, 4th Learner Corpus Research Conference, Bolzano, Italien, 05/10/2017.

Svensk ordbok 2009. Stockholm: Norstedts ordbok. https://svenska.se.

Utbildningsstyrelsen 2018. *Den gemensamma europeiska referensramen för språk: lärande, undervisning, bedömning. Kompletterande del och tilläggsdeskriptorer. Sammanfattning*. Utbildningsstyrelsen.

Utbildningsstyrelsen 2019. *Grunderna för gymnasiets läroplan 2019.* Utbildningsstyrelsen. Hämtad 20.10.2021. https://www.oph.fi/sites/default/files/documents/grunderna_for_gymnasiets_laroplan_2019.pdf.

Utbildningsstyrelsen 2021. *Nivåskalan för språkkunskap och språkutveckling*. Hämtad 19.10.2021. https://www.oph.fi/sv/utbildning-och-examina/nivaskalan-sprakkunskap-och-sprakutveckling.

Volodina, E., Alfter, D., Lindström Tiedemann, T., Piipponen, D. & Lauriala, M. 2021. Reliability of automatic linguistic annotation: native vs non-native texts. I: Monachini, M. & Eskevich, M. (red.), *CLARIN Annual Conference 2021. Proceedings*. Virtual edition. S. 90–94. https://office.clarin.eu/v/CE-2021-1923-CLARIN2021_ConferenceProceedings.pdf.

Volodina, E. & Johansson Kokkinakis, S. 2012a. Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish. I: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, 1040–1046.

Volodina, E. & Johansson Kokkinakis, S. 2012b. *Swedish KELLY: Technical report*. Forskningsrapporter från institutionen för svenska språket. GU-ISS-2012-01. http://hdl.handle.net/2077/28860.

Volodina, E., Pilán, I., Rødven Eide, S. & Heidarsson, H. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. I: *Proceedings of the third workshop on NLP for computer-assisted language learning*. S. 128–144.

Volodina, E., Pilán, I., Enström, I., Llozhi, L., Lundkvist, P., Sundberg, G. & Sandell, M. 2016a. SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies. I: *Proceedings of LREC 2016*, Slovenia.

Volodina, E., Pilán, I., Llozhi, L., Degryse, B. & François, T. 2016b. SweLLex: second language learners' productive vocabulary. I: *Proceedings of the workshop on NLP4CALL&LA*. NEALT Proceedings Series / Linköping Electronic Conference Proceedings.

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Wray, A. & Perkins, M. R. 2000. The functions of formulaic language: an integrated model. I: *Language and Communication* 20 (1). S. 1–28.