Automatisk målgruppsanpassad textanpassning

Evelina Rennes & Arne Jönsson
Institutionen för datavetenskap
Linköpings universitet, Linköping, Sverige

Språkteknologisk forskning gör det möjligt att utveckla teknik som automatiskt kan anpassa texter på olika sätt, exempelvis genom att sammanfatta eller att skriva om texter så att de blir enklare. Det är också möjligt att mäta texters komplexitet på olika nivåer genom att exempelvis mäta andelen svåra ord, meningskomplexitet och komplexitet för hela texten. Eftersom det inte finns en anpassad text som passar alla är det viktigt att göra teknikerna tillgängliga för såväl skribenter som läsare, så att de kan anpassa texten efter sin målgrupps behov. I artikeln presenterar vi ett antal tekniker för och utmaningar med detta. Vi illustrerar med exempel från system för mätning av textkomplexitet och olika typer av textanpassning.

1. Inledning

Att kunna läsa och förstå skriven text är i mångt och mycket en förutsättning för ett liv där man till fullo inkluderas i samhället. Detta gäller inte minst de senaste decennierna när den digitala utvecklingen har tagit ett jättesprång, och alltmer text produceras och görs tillgänglig på internet. Faktum är att läsning ses som en så pass viktig förmåga att det nämns i FN:s deklaration om mänskliga rättigheter. I *Konvention om rättigheter för personer med funktionsnedsättning*, Artikel 21, *Yttrandefrihet och åsiktsfrihet samt tillgång till information*, står det:

> Konventionsstaterna ska vidta alla ändamålsenliga åtgärder för att säkerställa att personer med funktionsnedsättning kan utöva yttrandefriheten och åsiktsfriheten, inklusive friheten att söka, ta emot och sprida uppgifter och idéer på lika villkor som andra och genom alla former av kommunikation som de själva väljer enligt definitionen i artikel 2, däribland genom att
> a) utan dröjsmål och extra kostnader förse personer med funktionsnedsättning med information som är avsedd för allmänheten i tillgängligt format och teknologi anpassad för olika funktionshinder, (SÖ 2008:26, s. 18)

Personer med olika typer av funktionsnedsättningar ska alltså ha möjlighet att kommunicera och ta till sig information på samma villkor som en person som inte har någon typ av läsproblem. Kort sagt: läsning är en mänsklig rättighet.

Även på nationell nivå är det sedan 2009 lagstadgat i språklagen (SFS 2009:600, § 11) att *"Språket i offentlig verksamhet ska vara vårdat, enkelt och begripligt"* och detta ställer naturligtvis stora krav på att myndighetstexter ska vara så begripliga som möjligt för att informationen ska nå så många som möjligt.

2. Lättläst och klarspråk

Lättläst och klarspråk är två begrepp som ofta, lite slarvigt, blandas ihop. På engelska är det än värre, där forskare och skribenter använder olika uttryck som Easy Language, Easy-to-Read, Easy Reading, Easy to Understand, Simplified Language, Clear Writing och Plain Language.

Klarspråk, det som på engelska ofta benämns som Plain Language, handlar om att myndighetstexter ska vara skrivna på ett *vårdat, enkelt och begripligt* språk (SFS 2009:600, § 11) för att uppnå målet att inkludera så många läsare som möjligt. Begreppet *lättläst* (MTM, 2021) innebär ett tydligare målgruppsfokus och inkluderar alla typer av texter. Lättlästa nyheter och lättlästa skönlitterära böcker är exempel på texter som faller under lättlästkategorin. I den här artikeln använder vi oss av båda begreppen, men även *lätt svenska* som här fungerar som ett paraplybegrepp för de båda termerna.

För att säkerställa att text skrivs på ett så enkelt och begripligt sätt som möjligt har det tagits fram riktlinjer för att skriva lättläst. I Sverige har Myndigheten för Tillgängliga Medier (MTM) tagit fram riktlinjer som innefattar både rent språkliga råd om ordval och grammatik, men också råd om bildsättning, textstruktur, radavstånd och typsnitt (Myndigheten för Tillgängliga Medier (MTM), 2021). Språkrådet har även gett ut en mer omfattande guide till klarspråk för myndigheter (Språkrådet, 2014).

Eftersom riktlinjerna för att skriva begriplig text försöker rikta sig till så många som möjligt kan de av naturliga skäl upplevas ganska generella och svepande. Denna breda ansats har givetvis fördelar, bland annat blir det enklare att träna upp nya skribenter om det bara finns en begränsad mängd enhetliga riktlinjer att förhålla sig till, men det finns även en del nackdelar. Den tyngsta är att det inte finns en text som passar alla läsare. Målgruppen för begripliga texter innefattar allt från personer med dyslexi, afasi, autism, kognitiv funktionsnedsättning till personer med annat modersmål än svenska, och dessa skilda målgrupper har väldigt olika behov och upplevda svårigheter. För att komplicera det ytterligare är det inte heller säkert att individer inom en målgrupp har samma behov.

För att kompensera för denna breda ansats tydliggörs det i olika riktlinjer att det är viktigt att ta reda på vem mottagaren är, och vilka behov mottagaren har, men i verkligheten kan texterna som skrivs ha flera möjliga mottagare. Finns det då tid och resurser för att anpassa texten till alla dessa mottagare?

Språkrådet utvärderar regelbundet klarspråksanvändningen hos svenska myndigheter, kommuner och regioner med hjälp av enkäter och intervjuer av klarspråksansvariga skribenter. I 2020 års version av denna utvärdering kom man bland annat fram till att det främsta hindret för att tillhandahålla information på klarspråk är tidsbrist (Hansson, 2020).

En möjlig lösning på problemet med tidsbrist är en effektiv teknisk lösning för att producera enkel text, en teknisk lösning som dessutom kan anpassa texterna efter personer med olika typer av läsbeteende och kanske till och med erbjuda en individuell anpassning.

3. Automatisk textanpassning för olika målgrupper

Att utveckla teknik för att automatiskt anpassa texter till målgrupper eller enskilda personer innebär en mängd utmaningar. Den största utmaningen handlar om att få grepp om läsaren. Har hen problem med figurativt språk, behöver hen få förklaringar på kulturella fenomen eller svåra ord? Har hen begränsat arbetsminne och behöver korta meningar med enkel syntaktisk struktur?

Vi vet en del om olika gruppers läsproblem. I tabell 1 ges en kortfattad sammanfattning av några typiska upplevda svårigheter hos olika grupper, men skillnaderna mellan olika individer är stora och den enskilde individen kan dessutom tillhöra fler än en målgrupp.

Tabell 1. Exempel på olika målgruppers svårigheter

Målgrupp	Exempel på upplevda svårigheter
Dyslexi	Långa och ovanliga ord, homofoner, ord som är ortografiskt lika, nya ord
Afasi	Hög informationsdensitet, långa meningar, långa sekvenser av adjektiv, passiv form, sammansatta ord
Andraspråksinlärning	Förståelse av ord och kulturella företeelser, textstruktur
Hörselskada	Komplexa grammatiska konstruktioner, textstruktur, långa meningar, förståelse av ord
Intellektuell funktionsnedsättning	Svårigheter relaterade till arbetsminnet, motivation till läsning, avkodning av text, textförståelse

Ytterligare utmaningar är att veta i vilken grad texter ska anpassas till respektive läsargrupp, i vilka sammanhang som anpassningar ska ske och med vilket

format de ska presenteras för läsaren. Dessutom är det viktigt att förenklingarna inte innebär att relevant information tas bort eller ändras. Det finns alltså en mängd delproblem som inte är triviala.

De språkteknologiska teknikerna för textanpassning utvecklas dock i rask takt, och verktyg som på olika sätt tar sig an problemet att öka den digitala inkluderingen görs i högre grad tillgängliga. Grovt kan de delas in i tre kategorier:

- **Tekniker för att mäta en texts komplexitet**. Dessa kan användas för att veta vilka egenskaper en text har och för att hitta texter som passar en viss grupp av läsare, eller till att förenkla texter med en viss grad av komplexitet.

- **Tekniker för att sammanfatta en text**. Här skiljer man vanligtvis mellan extraktiv och abstraktiv sammanfattning. I den extraktiva sammanfattningen plockas de viktigaste meningarna ut, i den abstraktiva sammanfattningen skrivs texten om helt.

- **Tekniker för att förenkla en text**. Detta kan bland annat ske syntaktiskt, genom att skriva om grammatiskt svåra meningar till enklare, eller lexikalt, där svåra ord och fraser byts ut mot enklare synonymer.

4. Textkomplexitetsmätning

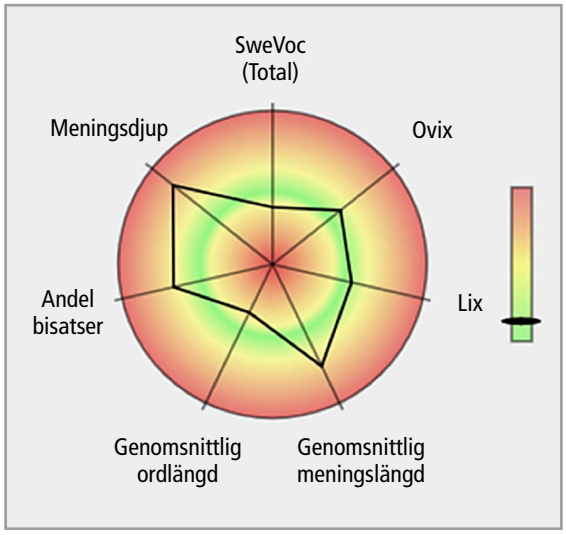
Om syftet är att anpassa en text så att den blir enklare behöver man först ta reda på hur svår texten är från början, och om det finns vissa komponenter i texten som är onödigt komplexa. Det kan göras med hjälp av en mer eller mindre sofistikerad komplexitetsmätning.

Dagens språkteknologiska tekniker låter oss nämligen att mäta det mesta i en text, och de mått som finns kan grovt delas upp i följande kategorier (Falkenjack et al., 2017):

- Ytliga mått
  - Mått som är enkla att räkna ut och inte behöver någon djupare analys, exempelvis antal ord per mening. Långa meningar är ofta svårare att förstå.
- Lexikala mått
  - Mått som är baserade på ordfrekvenser och grundläggande svensk vokabulär. Här ger ordlistorna från SweVoc (Heimann Mühlenbock & Johansson Kokkinakis, 2012) en bra grund för att ange vilka ord som är vanliga och vilka som är mindre vanliga. Ett vanligt och vardagligt ord upplevs oftare som enklare att läsa.

- Morfosyntaktiska mått
  - Mått som bygger på en morfologisk analys av texten och som gör att vi kan räkna exempelvis antalet innehållsord såsom substantiv, verb, adjektiv och adverb i en text. En hög andel indikerar en innehållsrik text som kan vara svår för individer med vissa typer av lässvårigheter. Å andra sidan kan en låg andel innehållsord indikera en syntaktiskt komplex text, vilket i sig kan utgöra en svårighet.
- Syntaktiska mått
  - Mått som speglar egenskaper beräknade efter en syntaktisk analys av texten. Hit hör bland annat dependenslängden, det vill säga avståndet mellan ett ord och dess huvud, oftast meningens verb. Texter med stora avstånd mellan sammanhängande ord är svårare att förstå än texter med mindre avstånd.
- Traditionella textkvalitetsmått
  - Mått som traditionellt har använts för att mäta läsbarhet, exempelvis LIX och OVIX.

Textkomplexitetsmåtten ger alltså en möjlighet att mäta en texts komplexitet utifrån flera perspektiv. För att måtten ska vara verkligt användbara krävs att de görs begripliga och går att koppla till individers upplevda läsning. Här ställs man inför två problem. Det första problemet är hur man ska tolka resultaten, som ofta kommer tillbaka som en tabell av siffror. En möjlighet skulle kunna vara att visualisera komplexiteten, exempelvis som i figur 1, i analogi med



Figur 1. Exempel på visualisering av textkomplexitet

Santini (Santini et al., 2020). Här behövs betydligt mer forskning kring vilka mått som bör visualiseras och även *hur* de bör visualiseras för att hjälpa en användare snarare än att stjälpa.

Det andra problemet är att koppla måtten till individers eller gruppers faktiska läsförmåga. Det vi ser i textkomplexitetsmåtten är själva textens egenskaper, som egentligen inte säger så mycket om hur en läsare kommer att uppleva läsningen av texten.

5. Automatisk textsammanfattning

Att erbjuda en läsare en sammanfattning kan underlätta genom att den viktigaste informationen i en text tydliggörs och samtidigt kan texten kännas kortare och enklare att ta till sig. Det finns två olika tekniker för att göra detta, extraktiv och abstraktiv sammanfattning. Extraktiv sammanfattning fungerar genom att de viktigaste meningarna plockas ut direkt från originaltexten och bildar en ny text, medan den abstraktiva sammanfattningen genererar en helt ny text som är en sammanfattning av originaltexten.

Antag att vi har följande, ganska långa, originaltext:

> Tillsammans med Kanadas nye stjärna Sidney Crosby duellerade 20-årige Alexander Ovetjkin om vem som skulle bli årets rookie i NHL. Poängmässigt vann den tuffe och spelskicklige ryssen i Capitals, laget från USA:s huvudstad. På 81 matcher under sin första säsong i NHL gjorde han 52 mål och passade till 54. 106 poäng en rookiesäsong är bland det absolut bästa en nykomling presterat i ligan. Han debuterade redan som 16-åring i den ryska ligan och trots locktoner om feta dollarbuntar från andra sidan av Atlanten stannade han kvar hemma i Moskva fram till och med förra säsongen. Det gjorde också att Ovetjkin var en redan klar elitspelare när han landade i Washington, där han för övrigt gjorde mål redan i sin debutmatch. Washington som lag räckte dock inte alls till för att ta sig till årets Stanley Cup-slutspel. Det är därför Alexander Ovetjkin nu kan komma och visa upp sig i Globen redan i kväll och därefter förstärka det ryska VM-laget i Riga. Det är alla fall var den ryska lagledningen hoppas på. Ovetjkin skulle egentligen ha anslutit till det ryska laget som på torsdagseftermiddagen kom från Helsingfors. Ryssland spelade sin första match i Hockey Games mot Finland i Helsingfors och vann i onsdags kväll. Nu blev Ovetjkin försenad från USA och landar inte i Stockholm förrän på fredagsmorgonen. Alexander Ovetjkin är bara en i raden av unga framgångsrika ryska hockeyspelare som kommer fram just nu. Tillsammans med bland andra Ilja Kovaltjuk, Atlanta och Jevgenij

> Malkin, Magnitogorsk i ryska ligan, har rysk hockey fått fram stjärnor som kan ta tillbaka landets hockeylandslag till en nivå vi inte sett sedan början av 1990-talet. Med Ovetjkin i laget blir det ännu svårare för Sverige att dels vinna dagens match i Globen och även att ta hem slutsegern i Euro Hockey Tour, som avgörs på måndag.

En extraktiv sammanfattning där man alltså plockar ut de viktigaste meningarna skulle då kunna se ut så här:

> Det är därför Alexander Ovetjkin nu kan komma och visa upp sig i Globen redan i kväll och därefter förstärka det ryska VM-laget i Riga. Alexander Ovetjkin är bara en i raden av unga framgångsrika ryska hockeyspelare som kommer fram just nu. Tillsammans med Kanadas nye stjärna Sidney Crosby duellerade 20-årige Alexander Ovetjkin om vem som skulle bli årets rookie i NHL. Med Ovetjkin i laget blir det ännu svårare för Sverige att dels vinna dagens match i Globen och även att ta hem slutsegern i Euro Hockey Tour, som avgörs på måndag.

Den extraktiva sammanfattaren har producerat en betydligt kortare text genom att plocka ut de viktigaste meningarna. Tekniken att göra detta bygger på att man skapar ordvektorer och representerar texten i ett flerdimensionellt vektorrum. Ordvektorer kan skapas på flera olika sätt men grundtanken är att ett ords betydelse beror av det sammanhang i vilket det används. Genom att till exempel tilldela varje ord en slumpmässig vektor av hög dimension, några hundra är vanligt, och sedan till varje ords vektor addera dess närmsta grannars ordvektorer kommer ordet att få en ny vektor som beror av det sammanhang i vilket ordet användes. Genom att upprepa detta tusentals gånger för varje ord i texter om miljontals ord kommer så småningom ordvektorerna att inte längre vara slumpmässiga. Ordvektormodeller finns fritt tillgängliga, för svenska t.ex. KB-BERT (Malmsten et al., 2020).

För att skapa en sammanfattning skapar man en vektorrepresentation för hela texten och varje mening för sig och sen väljer de meningar vars vektorer ligger närmast originaltexten. Eftersom orden är representerade som vektorer kan man använda cosinus för att direkt mäta hur nära meningarna ligger texten. Man tar också hänsyn till vilka koncept i texten som är viktigast och vilka meningar som innehåller detta, se exempelvis (Smith & Jönsson, 2011).

Den abstraktiva sammanfattaren, som tränats på en stor mängd redan sammanfattade texter, skriver i stället en sammanfattning med nya ord. Att hitta en stor mängd av korrekt sammanfattade texter som kan användas till träning

är svårt, men om man antar att ingressen till en nyhetsartikel sammanfattar brödtexten kan dessa användas som träningsdata (Monsen & Jönsson, 2021). I den abstraktiva sammanfattaren används återigen ordvektorer. Mer specifikt har modellen tränats för att känna igen vilken information i en given text som är relevant att ha med i sammanfattningen och, givet denna information, predicera vilket ord som följer på ett annat. När en ny text sedan matas in i modellen kommer den mest sannolika sammanfattningen, baserat på den tränade modellen, att matas ut. Ett exempel på detta, är följande:

> Alexander Ovetjkin är klar för spel i det ryska VM-laget i ishockey. Det innebär att han kommer att förstärka det ryska laget i kvällens match mot Sverige i Globen. Det blir Ovetjkins första match i Moskva och den blir tuffare än vad han hade tänkt sig innan han lämnade Washington Capitals för att spela i ryska ligan. Det är nu ryska hockeyligan, NHL, är på väg att ta upp sporten på den nivå man inte sett sedan början av 1990-talet. I dag möts Sverige och Ryssland i Globen i en match som kan bli avgörande för om Sverige ska vinna Euro Hockey Tour.

Här ser vi att meningarna är helt omskrivna, och modellen har skapat en helt ny text. Vi ser också att meningarna är grammatiskt korrekta men att innehållet skiljer sig något från det som genererats av den extraktiva sammanfattaren. Vilken sammanfattning som objektivt sett är bäst är inte lätt att säga.

Fördelarna med den extraktiva sammanfattningen är att den är relativt enkel, rent metodmässigt, och att sammanfattningen inte adderar någon ny information. Nackdelarna med den extraktiva sammanfattningen är att texten kan upplevas som lite stolpig och att relevant information kan gå förlorad i extraktionsprocessen. I exemplet har detta skett när första meningen inleds med ett ”därför” utan att det finns någon mening tidigare som motiverar detta ”därför”.

Den abstraktiva sammanfattningen har å andra sidan fördelen att det blir ett mer naturligt flyt genom texten, vilket kan förhöja läsbarheten och läsupplevelsen, men nackdelen att felaktig information riskerar att läggas till texten. I exemplet syns detta t.ex. i det felaktiga ”Det är nu ryska hockeyligan, NHL”.

Vi har i studier visat att den extraktiva sammanfattningstekniken i viss mån föredras om man presenteras för texter sammanfattade med hjälp av båda teknikerna samt originaltexten.

6. Automatisk textförenkling

Att göra en text syntaktiskt, lexikalt eller semantiskt enklare med hjälp av språkteknologiska verktyg kallas för automatisk textförenkling. Det kan bland

annat göras genom att byta ut ord i texten mot enklare synonymer, lägga till en förklaring till ett svårt begrepp, skriva om en mening så att den får en rakare ordföljd, eller ta bort onödiga inskjutna bisatser.

Automatiskt byte av synonymer görs relativt enkelt. Det finns språkliga resurser att tillgå på de flesta språk som möjliggör att ord kan slås upp och bytas ut mot andra. Däremot är det fortfarande en öppen fråga hur synonymerna ska väljas ut. Här kan studier av hur professionella skribenter gör för att förenkla sin text ge värdefull kunskap.

Synonymutbyten kan göras på olika sätt, exempelvis genom att byta ut ett ord baserat på frekvens, där vanligare ord ofta är enklare att förstå. Till exempel är *lön* (frekvens 1057 i den svenska Parole-ordlistan[1]) enklare att förstå för de flesta än *inkomst* (frekvens 378). Detta fungerar dock inte alltid, till exempel är *hustrun* (frekvens 608) betydligt vanligare än *frun* (frekvens 213). Då kan man i stället, eller som ett komplement, välja andra strategier för att hitta den mest passade synonymen. Sådana strategier kan vara att se till ordets längd och välja det kortaste ordet som synonym, eller att utgå från andra välstuderade egenskaper hos orden så som konkretionsgrad eller vid vilken ålder som ordet vanligen lärs in.

Synlex (Kann, 2004) är en resurs där föreslagna synonympar annoterats med grad av synonymitet, det vill säga, där personer har fått betygsätta hur synonyma två ord är. Men även här kan det bli svårt att välja rätt synonym. I Synlex listas till exempel *ersättning*, *gage*, *honorar* och *lön* som möjliga synonymer till *arvode* och med snarlika grader av synonymitet. Här kan sammanhanget möjligen hjälpa till.

Slutligen visar det sig att *basord*, eller vad som ofta kallas *prototyper*, ofta är enklare att förstå (Rennes & Jönsson, 2021), till exempel är *hund* enklare att förstå än *terrier* eller *dvärgschnauzer*. Men, precis som vid de andra strategierna för synonymbyte kan en del precision gå förlorad om man väljer basordet till förmån för det mer specifika ordet.

Automatisk omskrivning av komplicerade meningar till enklare kan göras på i princip två sätt, antingen genom att skriva regler som anger hur en mening kan förenklas, eller genom att träna modeller på parallellställda korpusar, i princip på samma sätt som maskinöversättning fungerar. Det finns för- och nackdelar med de olika metoderna. Regelbaserade system tillåter mer precisa och skräddarsydda omskrivningar, men det är tidskrävande att konstruera regler som fungerar i alla sammanhang. Modellbaserade system kräver inte mycket manuellt arbete, men däremot behövs en större mängd högkvalitativa

1 https://spraakbanken.gu.se/resurser/parolelex

och parallella data för att prestera ett bra slutresultat, vilket visar på vikten av data för textanpassning.

Regelbaserade textförenklingar bygger på att identifiera vissa språkliga konstruktioner i meningar och sedan skriva om dem. En regel som till exempel skriver om från passiv form till aktiv måste först göra subjektet i den passiva meningen till objekt i den aktiva. Därefter måste s-ändelsen på verbet tas bort och eventuell agent transformeras till subjekt. Meningen: *Den enorma kakan åts av både Kalle och Stina i matsalen* kommer då att skrivas om till: *Både Kalle och Stina åt den enorma kakan i matsalen*. Andra förenklingsregler kan handla om att tillämpa rak ordföljd (*Igår köpte Kalle en ny bil → Kalle köpte en ny bil igår*), att dela upp långa i flera kortare och citatomvandling (*”Gå och lägg dig!” sade Kalle → Kalle sade: ”Gå och lägg dig!”*).

Det finns förstås väldigt många möjliga förenklingsregler och en del fungerar genom att ta bort information som kanske kan vara onödig, till exempel vissa adjektiv (*Kalle har en stor blågrön bil → Kalle har en bil*). Problemet här är att inte ta bort för mycket som följande exempel från ett textförenklingssystem visar: *Europa är en av de mest urbaniserade kontinenterna i världen → Europa är en av kontinenterna*.

7. Språkliga data

Professionella skribenter har under många år anpassat texter till olika målgrupper. Sådana texter kan vi analysera och använda oss av för att få en förståelse av hur vi bör förenkla texter, vilket i sin tur kan användas till skrivstöd för skribenter eller för att utveckla de tekniker för automatisk textanpassning som beskrivits ovan.

För att utveckla olika tekniker för automatisk textanpassning kan korpusar av förenklade texter fungera som facit, en guldstandard, dit man vill nå med sina textförenklingar. Detta är användbart för att utvärdera regelbaserad teknik och är nödvändigt för modellbaserade tekniker som bygger på att man använder maskininlärning för att träna en modell på data som anses representera det man vill uppnå. Den abstraktiva sammanfattaren som presenterades ovan tränades till exempel på en korpus bestående av 38 151 artiklar från Dagens Nyheter där ingressen användes som sammanfattning av artikeln (Monsen & Jönsson, 2021).

Tanken bakom att använda textkorpusar på detta sätt är att det som de professionella skribenterna gör är korrekt och dit man vill sträva. Dessvärre finns det väldigt få texter på svenska som skrivits för en specifik målgrupp, där både de förenklade texterna och originaltexterna finns tillgängliga. I stället är korpusar på svenska med texter på lätt svenska hämtade från ett material som inte är målgruppsanpassat, eller som försöker nå en väldigt bred målgrupp.

För svenska finns det en större korpus av svenska lättlästa texter som heter LäSBarT[2]. LäSBarT innehåller drygt 1 miljon ord uppdelade på fyra olika genrer och har använts bland annat för att bättre förstå textförenkling och studera textkomplexitet (Falkenjack et al., 2013).

För att bygga modeller för textförenkling som beskrivits ovan vore det användbart att ha korpusar som är parallellställda, det vill säga där en mening har matchats ihop med en förenklad motsvarighet. En svensk sådan korpus innehåller samtliga svenska offentliga förvaltningars (myndigheter, kommuner och regioner) publika webbsidor från 2017 på vanlig svenska respektive lättläst svenska[3] (Rennes, 2018).

Tabell 2. Beskrivning av en svensk parallellställd korpus

	Vanlig svenska	Lätt svenska
Antal dokument	136 501	1629
Antal tecken	29,2 miljoner	334 491
Parallella	15 433 unika meningspar	

Korpusen beskrivs i tabell 2 där man tydligt ser att det finns väldigt få sidor på lätt svenska. Korpusen har också parallellställts genom att utgå från de enkla meningarna och leta efter meningar med motsvarande innehåll bland de texter som är skrivna på vanlig svenska (Rennes, 2020). För att räkna ut vilka meningar som har mest lika innehåll används återigen ordvektorer. Här tar man varje ordvektor i meningsparen, räknar ut hur lika de är, cosinusavståndet, och summerar detta för hela meningen. Därefter väljer man ut de meningspar vars summa överstiger en tröskel. Tröskelvärdet provas ut genom att be informanter avgöra hur lika två meningar är, en lättläst och en på vanlig svenska. Genom att göra detta för ett urval meningspar med olika cosinusavstånd får vi fram ett värde då meningsparen inte längre anses lika. Detta resulterade i 15 433 unika meningspar. Tyvärr är detta alldeles för få meningspar för att kunna träna en modell för automatisk textförenkling. Däremot kan den användas för att utveckla nya och bättre förenklingsregler.

8. Implementerade verktyg

Att kunna läsa och förstå skriven text är en förutsättning för att inkluderas i

2 https://spraakbanken.gu.se/resurser/lasbart

3 https://www.ida.liu.se/~arnjo82/diginclude/corpus.shtml (Rennes, An Aligned Resource of Swedish Complex-Simple Sentence Pairs, 2018)

samhället, och varje medborgare ska ha samma rätt att kunna ta till sig information. Vårt strå till stacken är en palett av tekniker och verktyg som kan användas både för att göra texter enklare att läsa och som ett skrivstöd för att skriva begripligt.

FriendlyReader är ett verktyg som riktar sig mot slutanvändaren. Idén med verktyget är att den enskilde läsaren själv ska kunna anpassa texten efter hens specifika behov. *FriendlyReader* innehåller i dagsläget möjligheten att sammanfatta texten, både extraktivt och abstraktivt, få synonymförslag till svåra ord och skriva om texten med enklare grammatisk struktur. Utöver de rent språkteknologiska funktionerna är det även möjligt att anpassa textstorlek, radavstånd, typsnitt och att få texten uppläst.

TeCST är ett verktyg som riktar sig mot skribenter och är tänkt som ett skrivstöd för att effektivt producera enkel text. *TeCST* innehåller möjligheten att få en texts komplexitet bedömd och visualiserad, att få förslag på syntaktiska förenklingar, samt få markeringar av svåra grammatiska strukturer och svåra ord i texten.

Vår intention med verktygen är att de ska kunna göra det möjligt att skräddarsy textanpassningar till en vald målgrupp, kanske till och med en enskild person. Bara genom att se till den enskilda läsaren och de utmaningar som denne ställs inför kan vi skapa texter som är tillgängliga på riktigt.

Summary

Language technology research makes it possible to develop technology for the automatic adaptation of text to enhance readability and comprehension. This can be done by creating automatic summaries of the text, or by rewriting the text so that it contains simpler words and less complex grammatical structures. It is also possible to measure the complexity of texts from various perspectives and text levels, including the proportion of difficult words, sentence complexity and measures on the complexity for the entire text. Since there is no *one-size-fits-all* text adaptation, it is important to make the techniques available to both writers and readers so that they can adapt the text to their needs and know how complicated it is. In the article, we present several techniques and illustrate with examples from real systems for text complexity measurement and text adaptation.

Författare

Evelina Rennes doktorerar i datavetenskap vid Linköpings universitet. I sin forskning intresserar hon sig framför allt för automatisk textanpassning för personer med olika typer av lässvårigheter.

Arne Jönsson är professor i datavetenskap vid Linköpings universitet. Han har under flera år forskat inom artificiell intelligens, de senaste åren med fokus på språkteknologi för ökad digital inkludering.

Litteraturförteckning

Falkenjack, Johan, Heimann Mühlenbock, Katarina & Jönsson, Arne, 2013: Features indicating readability in Swedish text. I: Stephan Oepen, Kristin Hagen, Janne Bondi Johannessen (red.): *Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa-2013), Oslo, Norge*. s. 27–40. Linköping: Linköping University Electronic Press.

Falkenjack, Johan o.a., 2017: Services for text simplification and analysis. I: Jörg Tidemann, Nina Tahmasebi (red.): *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa-2017) Göteborg, Sverige*, s. 309–313. Linköping: Linköping University Electronic Press.

Hansson, Karin, 2020: *"Det finns ett sug efter klarspråk" En studie om bättre stöd till klarspråk i offentlig verksamhet*. Stockholm: Institutet för språk och folkminnen. Tillgänglig på Internet: https://isof.diva-portal.org/smash/get/diva2:1440648/FULLTEXT01.pdf

Heimann Mühlenbock, Katarina & Johansson Kokkinakis, Sofie, 2012: SweVoc – A Swedish vocabulary resource for CALL. I: Lars Borin, Elena Volodina (red.): *Proceedings of the SLTC 2012 workshop on NLP for CALL, Lund, Sverige*, s. 28–34. Linköping: Linköping University Electronic Press.

Kann, Viggo, 2004: *Folkets användning av Lexin – en resurs*. Tillgänglig på Internet: http://www.csc.kth.se/~viggo/rapporter/synlex.pdf

Malmsten, Martin, Börjeson, Love & Haffenden Chris, 2020: Playing with Words at the National Library of Sweden – Making a Swedish BERT, *arxiv.org:2007.01658v1 [cs.CL].*

Mühlenbock, Katarina, 2008: Readable, Legible or Plain Words – Presentation of an easy-to-read Swedish corpus. I: Anju Saxena, Åke Viberg (red.): *Multilingualism: Proceedings of the 23rd Scandinavian Conference of Linguistics*, s. 327–329. Uppsala: Acta Universitatis Upsaliensis.

Monsen, Julius & Jönsson, Arne, 2021: A method for building non-English corpora for abstractive text summarization. I: Monica Monachini, Maria Eskevich (red.): *Proceedings of the CLARIN annual conference*, s. 82–85.

Myndigheten för Tillgängliga Medier (MTM), 2021: *Att skriva lättläst*. Tillgänglig på Internet: https://www.mtm.se/var-verksamhet/lattlast/att-skriva-lattlast/

Rennes, Evelina, 2018: An Aligned Resource of Swedish Complex-Simple Sentence Pairs. I: *Proceedings of the Seventh Swedish Language Technology Conference (SLTC), Stockholm, Sverige*, s. 61–63.

Rennes, Evelina, 2020: Is it simpler? An Evaluation of an Aligned Corpus of Standard-Simple Sentences. I: Núria Gala, Rodrigo Wilkens (red.): *Proceedings of the LREC Workshop on Tools and Resources to Empower People with REAding Difficulties (READI-20), Marseille, France*, s. 6–13. European Language Resources Association.

Rennes, Evelina & Jönsson, Arne, 2021: Synonym Replacement based on a Study of Basic-level Nouns in Swedish Texts of Different Complexity. I: Simon Dobnik, Lilja Øvrelid (red.): *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), Reykjavik, Island*, s. 259–267. Linköping: Linköping University Electronic Press.

Santini, Marina, Jönsson, Arne & Rennes, Evelina, 2020: Visualizing Facets of Text Complexity across Registers. I: Núria Gala, Rodrigo Wilkens (red.): *Proceedings of the LREC workshop Tools and Resources to Empower People with REAding DIfficulties (READI-20)*, s. 49–56. European Language Resources Association.

Smith, Christian & Jönsson, Arne, 2011: Enhancing extraction based summarization with outside word space. I: Haifeng Wang, David Yarowsky (red.): *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP), Chiang Mai, Thailand*, s. 1062–1070. Asian Federation of Natural Language Processing.

Språkrådet, 2014: *Myndigheternas skrivregler, 8 upplagan*. Stockholm: Norstedts Juridik AB/Fritzes. Tillgänglig på Internet: https://www.isof.se/download/18.17dda5f1791cdbd2873a99/1620030264840/Mynd-skrivreg2014-1.pdf

SÖ 2008:26, 2008: *Konvention om rättigheter för personer med funktionsnedsättning*. Tillgänglig på internet: https://www.regeringen.se/4ae1cb/globalassets/regeringen/dokument/socialdepartementet/funktionshinder/konvention-om-rattigheter-for-personer-med-funktionsnedsattning.pdf

Nyckelord

automatisk textanpassning, textkomplexitet, automatisk textsammanfattning, automatisk textförenkling, lässvårigheter