Förenklad textinmatning på mobila enheter med hjälp av kontextbaserad språktolkning

Examensarbete utfört i Informationsteori

av

Anders Jensen

LITH-ISY-EX--05/3607--SE

Linköping 2005
Förenklad textinmatning på mobila enheter med hjälp av kontextbaserad språktolkning

Examensarbete utfört i Informationsteori vid Linköpings tekniska högskola av
Anders Jensen

LITH-ISY-EX--05/3607--SE
Förenklad textinmatning på mobila enheter med hjälp av kontextbaserad språktolkning

Simplified text input for mobile devices using context based language interpretation

Anders Jensen

Förenklad textinmatning på mobila enheter med hjälp av kontextbaserad språktolkning

Simplified text input for mobile devices using context based language interpretation

Förenklad textinmatning på mobila enheter med hjälp av kontextbaserad språktolkning

Simplified text input for mobile devices using context based language interpretation

Nyckelord

textinmatning, sms, mobiltelefon, kontextmodellering, textmeddelande, begränsad knappsats, språkmåling, viterbiavkodning

Keyword

textinput, sms, mobile phone, context modeling, text message, limited keypad, language modeling, viterbi decoding
Abstract

The number of text messages sent from mobile phones, has increased dramatically over the last few years. Along with that, we are witnessing a lot of new mobile portal services currently being developed. Many of these services rely on an ability to input text efficiently. The traditional phone keypad is ambiguous because each key encodes more than one letter. At present, the most common way to deal with this problem is using a stored dictionary to guess the intended input.

This thesis presents a new text entry strategy called Qtap. Instead of using a stored dictionary to guess the intended word, this method uses probabilities of letter sequences. New features that come with Qtap are the usage of the viterbi algorithm to decode input sequences and a non-alphabetic keypad. How the strategy and the keypad used by Qtap were developed, is described throughout the thesis.

Qtap is also compared to a dictionary-based method, t9, on a non-user level. The results show Qtap is performing well in many senses. The conclusion from this is that a further development of Qtap is motivated.

A discussion of various modifications and additions to the design, that may yield a performance improvement, is also included.

Sammanfattning


Qtap jämförs också med det ordlistebaserade systemet t9. Jämförelsen sker på en nivå som inte inkluderar interaktion mellan system och användare. Qtap uppvisar goda testresultat i många avseenden och slutsatsen som dras från detta är att en vidare undersökning av Qtap är motiverad.

Rapporten innehåller också en diskussion kring tänkbara modifikationer och tillägg som kan ge förbättrad systemprestanda.
Innehåll

ABSTRACT __________________________________________________________ VII
SAMMANFATTNING __________________________________________________ VII
INNEHÅLL __________________________________________________________ IX
1 INLEDNING ________________________________________________________ 1
  1.1 BAKGRUND ____________________________ 1
  1.2 Mål _________________________________ 2
  1.3 METOD ______________________________ 2
  1.4 AVGRÄNSNINGAR ______________________ 2
  1.5 DISPOSITION _________________________ 3
2 DAGENS SYSTEM __________________________________________________ 4
  2.1 KNAPPSATSEN _________________________ 4
  2.2 MULTITAP ___________________________ 5
    2.2.1 Bakgrund ____________________________ 5
    2.2.2 Utvärdering _________________________ 5
  2.3 SYSTEM BASERADE PÅ ORDLISTA __________ 5
    2.3.1 Bakgrund ____________________________ 5
    2.3.2 Utvärdering _________________________ 6
  2.4 LETTERWISE __________________________ 7
    2.4.1 Bakgrund ____________________________ 7
    2.4.2 Utvärdering _________________________ 7
  2.5 FRAMTIDSUTSIKTER ______________________ 8
3 SPRÅKSTATISTIK ________________________________________________ 9
  3.1 BAKGRUND ____________________________ 9
  3.2 BEGRÄNSNINGAR ______________________ 9
  3.3 INSAMLAT MATERIAL ____________________ 10
    3.3.1 Bokstavsfrekvenser ____________________ 10
    3.3.2 Bigram ______________________________ 11
    3.3.3 Trigram _____________________________ 11
    3.3.4 Tetragram ____________________________ 11
4 TEORI ____________________________________________ 12
  4.1 BAKGRUND ____________________________ 12
  4.2 MARKOVMODELLER ______________________ 13
  4.3 DOLDA MARKOVMODELLER ________________ 15
    4.3.1 Exempel ____________________________ 17
    4.3.2 Terminologi __________________________ 18
    4.3.3 Spaljédiagram ________________________ 18
  4.4 VITERBIAVKODNING ______________________ 20
    4.4.1 Viterbialgoritmen ______________________ 20
    4.4.2 Exempel ______________________________ 21
5 DET NYA SYSTEMET ____________________________________________ 25
  5.1 FÖRUTSÄTTNINGAR _____________________________________________ 25
    5.1.2 Utforande ____________________________ 25
    5.1.3 Testtext ______________________________ 25
1 Inledning

1.1 Bakgrund

Antalet sända textmeddelanden via mobiltelefoner har stigit dramatiskt de senaste åren och i Sverige översteg siffran 2 miljarder år 2004. Samtidigt ökar mobiltelefontätheten i världen lavinartat och det finns inte något som tyder på att trenden kommer att avta under överskådlig framtid. Utbudet av mobila tjänster som kräver textinmatning har också blivit större och allt fler utnyttjar idag mobiltelefonen för att sända och ta emot e-post, söka information på Internet, föra minnesanteckningar och liknande. Allt detta ställer självklart högre och högre krav på mobiltelefonens textinmatningsystem vad gäller effektivitet och användarvänlighet.

Det begränsade antalet knappar på en traditionell mobiltelefon sätter emellertid gränser för vad som kan uppnås i dessa båda avseenden. Flera symboler måste placeras på samma knapp, vilket medför en tvetydighet vid textinmatning och frågan är hur denna tvetydighet ska elimineras. På dagens mobiltelefoner finns det väsentligen två olika strategier för att lösa detta bekymmer.

Den första, som kallas multitap, går ut på att användaren genom upprepade knapptryckningar ger systemet direkt information om önskad symbol och på så sätt eliminerar tvetydigheten.

Den andra strategin utnyttjar en lagrad ordlista för att gissa vilken symbol som avses vid knapptryckning. I Europa är t9 från Tegic Communications det överlägset mest spridda exemplet på ett system som bygger på denna teknik.[1]

Inget av dessa båda tillvägagångssätt bedöms emellertid som helt tillfredsställande av genomsnittsanvändaren. Multitap är väldigt tidskrävande och kräver ett stort antal knapptryckningar, något som givetvis rimmer illa med informationssamhällets höga krav på effektivitet. T9 är visserligen är en klar förbättring jämfört med multitap men då systemet gissar på fel ord eller det önskade ordet saknas i ordlistan, uppstår likväl förhållanden som leder till irritation hos många användare.

1.2 Mål

Mål med detta examensarbete är att utarbeta och undersöka ett system för textinmatning på mobiltelefon som bygger på sannolikheter för olika mönster i det svenska språket. I sökandet efter lämplig strategi ska olika sätt att utnyttja de språkstrukturer de sannolikheterna betraktas och jämföras med varandra. I målet ingår också att undersöka om en förändring av symbolernas placering på knappsatsen kan förbättra systemet ytterligare. Till sist ska det framtagna systemet utvärderas mot t9.

1.3 Metod

Vid designen av det nya systemet har ett primitivt system valts som utgångspunkt. Därefter har stegvisa förbättringar av systemet gjorts tills tillfredsställande resultat erhållits. Samtliga ansatser har implementerats och utvärderats med hjälp av Matlab som också har använts för att generera den statistik som ligger till grund för hela arbetet. Statistik över t9’s prestanda har framtagits genom manuell inmatning på en mobiltelefon av modell Sony Ericsson T610.

1.4 Avgränsningar

Ämnets omfattning gör vissa avgränsningar nödvändiga. De krav som systemet ställer på minneskapacitet och frågan om lämpligt användargränssnitt behandlas ej utöver några korta kommentarer i den avslutande diskussionen.

Att användargränssnitt inte har studerats får konsekvenser vid utvärderingar av olika system. Vid samtliga jämförelser har bara grundsystemen i sig betraktats och de eventuella fördelar ett system kan uppnå via interaktion med användaren tas inte med i beräkningarna.

Genom hela arbetet har endast det svenska språket studerats. Stora delar av de resultat och diskussioner som presenteras kan dock lätt generaliseras till andra språk. Inga symboler utöver det svenska alfabetets 29 bokstäver samt blanksteg har betraktats och allt statistikmaterial och samtliga texter har därför anpassats för att uppfylla detta krav. Detta beror bland annat på att allt för många specialfall inte är önskvärt, men också på att det inte finns någon mobiltelefonstandard för inmatning av avvikande symboler.

Slutligen har undersökningen av alternativa knappsatsar endast involverat omplacering av symbolerna och inte utnyttjat möjligheten att ändra den fysiska utformningen genom att ta bort eller lägga till knappr.

Begränsningarna beskrivs ofta i detalj i anslutning till de kapitel de berör.
1.5 Disposition

Rapporten är utformad enligt följande:

Kapitel 2 redogör för hur dagens system för textinmatning fungerar. En närmare beskrivning av multitap, t9 och Letterwise ges.

Kapitel 3 behandlar det statistikmaterial som skall ligga till grund för det nya systemet. Förutom en beskrivning av vilka data som har samlats in finns också information om det textmaterial som all statistik baserats på.

Kapitel 4 beskriver i allmänna ordalag en del bakomliggande teori som kommer att utnyttjas längre fram.

Kapitel 5 beskriver i kronologisk ordning de olika systemansatser som har undersöks och så småningom mynnat ut i det slutliga valet av ett system för vidare undersökning.

Kapitel 6 undersöker om det system som valdes i kapitel 5 går att förbättra ytterligare genom att variera knappsatsens utseende.

Kapitel 7 är ett statistikavsnitt som avser att jämföra den nya strategin med t9 ur olika aspekter. För- och nackdelar med de olika angreppssätten påvisas via olika undersökningar.

Kapitel 8 innehåller huvudsakligen slutsatser och en diskussion kring vilka förutsättningar som råder för en fortsatt undersökning av systemet. Dessutom finns en kortare diskussion kring framtiden för textinmatning på mobila enheter.

Appendix A omfattar en del av den språkstatistik som legat till grund för arbetet.

Appendix B innehåller tabeller som rör det mått på språklig funktion som diskuteras i kapitel 6.

Appendix C innehåller några av de testtexter som används i kapitel 6.
2 Dagens system

Det här kapitlet beskriver dagens kommersiella system för textinmatning på mobiltelefon. Förutom en beskrivning av funktionen hos varje system ges en kort utvärdering med systemens för- och nackdelar. Slutligen berörs de olika strategiernas utvecklingsmöjligheter och valet att studera sannolikhetsbaserade system motiveras.

2.1 Knappsatsen

Traditionella mobiltelefoner är utrustade med en knappsats som omfattar 12 knappar för enkel inmatning. Vanligvis är knapparna 2-9 ägnade åt alfabetets bokstäver medan blanksteg återfinns på knapp 0 eller knapp #. Designdetailjer kan dock skilja sig åt mellan olika tillverkare och en enhetlig svensk standard existerar inte än så länge. Den knappsats vi ska utgå ifrån och använda genom hela arbetet då inte annat anges, är den som återfinns i figur 2.1. Blanksteg har här placerats på knapp 0 och kommer i fortsättningen betecknas med _.

Eftersom det genomgående är det svenska språket som studeras spelar också placeringen av Å, Ä och Ö en avgörande roll. Vi kommer att placera Å och Ä på knapp 2 och Ö på knapp 6 (denna placering är den vanligaste men det kan skilja sig mellan olika tillverkare och modeller). Som vi ska se längre fram kommer vi inte använda oss av andra knappar än de nio som är reserverade för bokstäver och blanksteg.

![Figur 2.1: Den traditionella knappsatsen.](image-url)
2.2 Multitap

2.2.1 Bakgrund

Den ursprungliga tekniken för att mata in text på en mobiltelefon kallas multitap. Vid multitap trycker användaren ned varje knapp ett antal gånger som motsvarar det önskade tecknet. På en traditionell knappsats innebär det till exempel att knapp 2 trycks ned en gång för A, två gånger för B och tre gånger för C. Denna teknik används fortfarande i hög utsträckning för vissa ändamål (till exempel inmatning i telefonbok) och är också den teknik man vanligtvis tillgriper då ”smartare” system misslyckas. Misslyckande kan till exempel bestå i att ett system baserat på lagrad ordlista inte finner det ord anvärdaren önskar skriva.

Vid multitap måste användaren också vänta tills en så kallad timeout inträffar innan nästa tecken kan matas in. För att mata in ordet BAD måste användaren alltså trycka två gånger på knapp 2, vänta en viss kritisk tid, trycka ytterligare en gång på knapp 2 för att därefter trycka en gång på knapp 3. Alternativet till att vänta mellan inmatning av B och A är att trycka på en speciell timeoutknapp som finns på i stort sett alla mobiltelefoner och på så vis tvinga fram en timeout. Mellan inmatningen av A och D krävs inte detta då byte av knapp automatiskt framkallar en timeout.

2.2.2 Utvärdering

Behovet av ett timeoutsystem och det stora antal knapptryckningar som krävs vid inmatning är de främsta nackdelarna med multitap. Fördelarna är att det inte finns någon som helst tvetydighet vid inmatningen. Utdata motsvarar alltid precis det användaren vill skriva.

2.3 System baserade på ordlista

2.3.1 Bakgrund

Den i dagsläget mest spridda tekniken för att underlätta textinmatning på mobiltelefon och råda bot på de problem som medföljer multitap, är ordlistebaserade system. Kommersiella produkter som utnyttjar detta inkluderar t9 från Tegic Communications[1], eZiText från Zi Corporation[3] och iTAP från Motorola[4].

När ett ordlistebaserat system nyttjas trycker användaren endast en gång på varje knapp. Vid inmatning av ordet TIO trycker vi således in knappsekvensen 246, vilket ger 3·3·3 = 27 möjliga bokstavskombinationer. Systemet jämför då dessa med en lagrad frekvensordlista och skriver ut det mest troliga - i det här fallet just det önskade ordet TIO. Jämfört med multitap är detta en avsevärd förbättring eftersom vi annars skulle ha tvingats trycka sekvensen 24444666 för att få samma ord.

Ett annat uppenbart bekymmer med dessa system är att ordlistan är begränsad och inte kan innehålla alla de ord man vill skriva. Det vanligaste sättet att komma runt detta är att användaren genom en knapptryckning stänger av det ordlistebaserade systemet och matar in det önskade ordet med hjälp av multitap. Detta ökar antalet knapptryckningar väsentligt, inte enbart på grund av multitaps ineffektivitet, utan också för att användaren ofta först försöker skriva in sitt ord med det ordlistebaserade systemet inkopplat och tvingas radera och göra om då det visar sig att ordet inte ingick i ordlistan. Ett annat sätt att komma runt problemet är komplettera ordlistan med egna ord vilket är möjligt på de flesta moderna mobiltelefoner.

2.3.2 Utvärdering

System som baseras på ordlista fungerar i regel mycket bra. Speciellt användare som vant sig vid den långsamma inmatningen med multitap upplever t9 och liknande system som en avsevärd förbättring. Att det ord man vill mata in inte alltid är det vanligaste brukar inte upplevas som en betydande brist då bläddringsfunktionen ofta är smidig och lättillgänglig. Ordlistans begränsning leder dock till att systemet på sikt kan uppfattas som svåröverskådligt. Ibland kan till synes självklara ord saknas i ordlistan medan andra, betydligt mer udda ord, finns med. Förutom att detta kan skapa irritation leder det också till att det på förhand är svårt att veta när systemet bör kopplas ur, även för erfarna användare. Möjligheten att lägga till egna ord i ordlistan är förvisso bra men underlättar främst i de fall ordlistan saknar ett ord användaren skriver ofta.

En annan faktor som måste beaktas vid en utvärdering av ordlistebaserade system är vad för slags text som matas in. Idag står ungdomar för en stor del av de textmeddelande som sänds och språket präglas därefter. Moderna uttryck, slangord, udda akronymer och ovanliga egennamn är frekvent förekommande och täcks inte på samma sätt av den lagrade ordlistan som bygger på mer traditionell svensk text.

Den stora fördelen med t9 och liknande system är givetvis den uppenbara förbättringen jämfört med multitap vad gäller snabbhet och antal knapptryckningar. Som påpekades tidigare har även vana användare svårt att på förhand ha en uppfattning om ett ovanligt ord förekommer i ordlistan eller ej. Däremot har personer med lång erfarenhet av ordlistebaserade system ofta en känsla för hur systemet reagerar vid inmatning av vanligare ord. Detta underlättar och ökar snabbheten på inmatningen då inte visuell kontroll krävs i samma utsträckning. Användaren behöver inte stanna upp
2.4 Letterwise

2.4.1 Bakgrund

Det finns i dagsläget endast en kommersiell produkt som istället för en lagrad ordlista utnyttjar kännedom om språkets struktur. Systemet går under namnet Letterwise och har utvecklats av Eatoni ergonomics[2].


Letterwise betraktar bara ett ord i taget. Detta innebär att det i början av varje ord inte finns något prefix att bygga förutsägelsen av utdata på, varför mer primitiva strategier får tillgripas. Första bokstaven i ett ord avgörs därför via ren bokstavsfrekvens och andra bokstaven baseras på en frekvenstabell över alla kombinationer av två bokstäver. Detta avser givetvis en begränsning i systemet eftersom användaren inte alltid önskar börja ord med den bokstav som är mest sannolik. För att råda bot på detta får användaren vid varje inmatning chans att korrigera valet av utdata. Detta säkerställer att de två föregående tecknen som så småningom studeras vid varje ny knapptryckning, verkligen motsvarar det användaren vill skriva. Hade detta inte varit fallet hade alla förutsägelser riskerat att bygga på felaktiga förutsättningar, vilket givetvis skulle vara fatalt.

2.4.2 Utvärdering

En bra egenskap hos den strategi som Letterwise utnyttjar är att den möjliggör inmatning av de ord som vi tidigare konstaterade att ordlistebaserade system har problem med. Såvida inte det vi försöker mata in på avsevärt sätt avviker från svensk språkstruktur har vi goda chanser att få korrekt utdata utan allt för många korrigerings.

2.5 Framtidsutsikter

Som vi har sett har alla kommersiella system sina för- och nackdelar. Frågan man kan ställa sig är vilka förutsättningar det finns för att förbättra de olika strategierna. Vi kan inte förvänta oss att komma särskilt mycket längre med multitap. Frågan blir i själva verket trivial eftersom definitionen av multitap inte tillåter några förändringar av systemet som sådant. Den enda möjligheten för att minska antalet knapptryckningar är därför att utöka antalet knappar på mobiltelefonen. En undersökning av denna möjlighet fäller dock utanför ramen för detta arbete. (Kapitel 8 innehåller dock en kortare diskussion kring utseendet hos framtidens knappsatser).


Mot bakgrund av detta kan vi hävda att system som bygger på språkliga sannolikheter är de mest lovande. Detta är också ett av huvudskälen till att vi kommer att studera just sådana system närmare. Innan undersökningen av sannolikhetsbaserade system kan ta sin början krävs emellertid en insamling av ganska omfattande statistikmaterial, vilket är föremål för nästa kapitel.
3 Språkstatistik

I det här kapitlet beskrivs det textmaterial som legat till grund för all den språkstatistik som utnyttjats i arbetet. Dessutom ges kommentarer till de olika sorters statistik som insamlades. Mer utförliga statistiktabeller finns i appendix A.

3.1 Bakgrund

En förutsättning för studiet av textinmatningssystem som bygger på sannolikheter, är tillgång till pålitlig statistik över frekvenser och bokstavsmönster i det svenska språket. Statistik över ett språk kan tas fram på många olika sätt och för att vara förvissad om att de data som utnyttjats är genererade på ett enhetligt sätt samlar vi in all statistik på egen hand. Utgångspunkten är ett textmaterial som går under namnet PAROLE och är insamlat inom ramen för ett EU-projekt med samma namn[5]. Den svenska delen av arbetet har utförts av Institutionen för svenska vid Göteborgs universitet under ledning av Daniel Ridings och består av drygt 19 miljoner löpande ord. Textmaterialet är hämtat från källor med skilda typer av text som romaner, dagstidningar och tidsskrifter för att utgöra en så bra representation som möjligt av det svenska språket.

Vi kommer att göra det rimliga antagandet att inte hela textmängden behöver utnyttjas för att få tillräckligt god statistik för våra ändamål. Vi reducerar därför materialet till drygt två miljoner ord bestående av knappt tio miljoner tecken (blanksteg ej medräknade). Samtliga skiljetecken och ord som innehöll tecken som inte ingår i det svenska grundalfabetet, till exempel É, Ü och Ñ, kommer vi att utesluta.

I den fortsatta texten kommer namnet PAROLE alltid att syfta på det reducerade materialet.

3.2 Begränsningar

3.3 Insamlat material

3.3.1 Bokstavsfrekvenser


![Bokstavsfrekvenser](image)

Figur 3.1: Bokstavsfrekvenser för Mühlenbocks sammanställning och vårt eget textmaterial, PAROLE.

Vi kan konstatera att skillnaderna mellan frekvenserna i de båda textmaterialen är små eller mycket små vilket, givet vår tilltro till Mühlenbocks material, ger oss anledning att betro vår egen metod för statistikinsamling. En intressant detalj, om än kuriös för våra ändamål, är att E är den mest förekommande bokstaven i Mühlenbocks material medan A är
Förenklad textinmatning på mobila enheter med hjälp av kontextbaserad språktolkning

frekventast i PAROLE. Vilken som är den vanligaste bokstaven i svensk text tycks alltså vara en frågeställning som fortfarande lever.

Till våra framtida ändamål behöver vi inte bara statistik över bokstavsfrekvenser. Vi vill också ha tillgång till liknande frekvenstabeller över så kallade bigram, trigram och tetragram i svenska. Ett bigram är en kombination av två bokstäver, till exempel GU och JL. Trigram och tetragram är på samma sätt en kombination av tre respektive fyra bokstäver.

3.3.2. Bigram

Antalet möjliga bigram med de 30 tecken vi betraktar är $30^2 = 900$. Av dessa förekommer 806 minst en gång i PAROLE och direkt kan vi alltså klassa 94 bigram som väldigt osannolika vid normal textinmatning (till exempel BZ, HQ och WJ).

Utöver de 94 som inte förekommer någon gång är det ytterligare 87 bigram som har färre än 10 belägg. Det allra vanligaste bigrammet i den reducerade versionen av PAROLE är NA som figurerur hela 327559 gånger.

3.3.3 Trigram

Analogt med vår tidigare insamling får vi antalet möjliga trigram till $30^3 = 27000$. Endast 11678 av dessa förekommer i vårt statistikmaterial och studerar vi bara de som förekommer minst 10 gånger är siffran 7891. Det vanligaste trigrammet i PAROLE är EN_ med 158512 belägg. (De bigram som bygger upp detta trigram, N_ och EN hade 327560 respektive 205535 belägg).

Trigramstatistiken kommer vi ha anledning att återkomma till ett flertal gånger framöver. Vi kommer att utnyttja trigramfrekvenser både i vår jakt på ett bra inmatningssystem och i våra försök att hitta en bättre knappsats. Vi påminner också om att det är just sådan här statistik över trigramfrekvenser som ligger till grund för Letterwise.

3.3.4 Tetragram

Frekvenser för längre bokstavskombinationer än tetragram kommer vi inte att utnyttja i detta examensarbete. Antalet tetragram blir $30^4 = 810000$ men endast 44716 av dessa förekommer i PAROLE. Vanligast bland dessa är _OCH med 12888 belägg.
4 Teori

Detta kapitel inleds med en kort fördjupning av textinmatningsproblemet. Därefter beskrivs en del grundläggande teori bakom *dolda markovmodeller* och *viterbiavkodning*[6],[7]. Denna teori kommer vi att utnyttja längre fram när vi söker efter ett nytt system.

4.1 Bakgrund


Vid varje knapptryckning uppstår nya möjligheter och antalet tänkbbara utdata växer snabbt. För att visualisera inmatningen kan vi använda oss av en diagramstruktur. Inmatningen av ordet REN skulle till exempel motsvaras av följande diagram:

![Diagram av inmatningen av ordet REN](image)

*Figur 4.1: Visualisering av inmatningen av ordet REN.*
Alla vägar vi kan ta genom diagrammet motsvarar en möjlig utdatasekvens. I det här fallet önskar alltså användaren mata in REN och det kommer vara vår uppgift att förutsäga detta. Till vår hjälp har vi endast vetskapen om vad användaren tryckte in, d.v.s. knappsekvensen 7 5 6.


En tillståndssbeskrivning likt den vi just beskrivit skulle ge oss en bra modell av vårt avkodningsproblem. Det skulle vara enkelt att på olika sätt utöka modellen för att få bättre förutsägelser. Vi förhoppning är dessutom att slutligen finna en praktisk metod för att avgöra vilken av alla möjliga vägar genom diagrammet som är den mest sannolika.

Som vi snart skall se kan dessa önskemål uppfyllas med hjälp av så kallade dolda markovmodeller. Dessa visar sig vara väldigt lämpliga för att modellera vårt problem. Att utgå ifrån markovmodeller gör också att det så småningom blir väldigt naturligt att hitta den optimala tillståndsekvensen, genom att utnyttja en strategi som kallas viterbiavkodning.

### 4.2 Markovmodeller


Vi utgår från att ”väderförutsägelser” i det här fallet går ut på att gissa morgondagens väder utifrån observationer av hur väderet har bett sig tidigare. För att kunna sätta upp en statistisk modell för väderförutsägelser samlar vi in statistik som beskriver hur väderet $q_n$ är idag (på dag $n$) givet väderförhållandet igår $q_{n-1}$, dagen dessförinnan $q_{n-2}$, o.s.v. Vi strävar alltså efter att finna följande villkorliga sannolikhet:

$$P(q_n \mid q_{n-1}, q_{n-2}, \ldots, q_1)$$

Uttryck 1 är alltså ett mått på sannolikheten för förutsägelsen av morgondagens väder, givet föregående dagars väderförhållanden. Om vi till exempel känner till att väderet de tre senaste dagarna var [$☀, ☀, ❄$] i kronologisk ordning, ges sannolikheten för att kommande dag är regnig av:

$$P(q_4 = ⛈ \mid q_3 = ❄, q_2 = ☀, q_1 = ☀)$$
Förenklad textinmatning på mobila enheter med hjälp av kontextbaserad språktolkning

Ett problem med denna strategi är att den informationsmängd om tidigare vädermönster vi måste samla in, växer väldigt snabbt då n växer. Detta har föranlett att man ofta gör en förenkling, kallad Markovs antagande, av ekvation 1:

För en sekvens \([q_1, q_2, \ldots, q_n]\) gäller:

\[
P(q_n | q_{n-1}, q_{n-2}, \ldots, q_1) = P(q_n | q_{n-1})
\]

(2)

Denna ekvation beskriver en så kallad markovmodell av första ordningen. I en sådan modell säger vi alltså att en viss observation vid tidpunkten \(n\) bara beror på observationen \(q_{n-1}\) vid tidpunkten \(n-1\). (I en markovmodell av andra ordningen skulle sannolikheten för en observation bero på två tidigare observationer, o.s.v.) Vi kan nu enkelt uttrycka den totala sannolikheten för en viss sekvens \([q_1, q_2, \ldots, q_n]\), kallad markovkedja, med hjälp av förenklingen i ekvation 2:

\[
P(q_1, \ldots, q_n) = \prod_{i=2}^{n} P(q_i | q_{i-1})
\]

(3)

Låt oss nu godtyckligt välja sannolikhet för morgondagens väder givet dagens väderförhållande, dvs:

\[
P(q_{	ext{imorgon}} | q_{	ext{dag}})
\]

Värdena vi har valt syns i figur 4.2. Vi kan välja i princip vilka värden som helst men eftersom det under alla omständigheter kommer att bli något slags väder imorgon så måste tabellens rader summera till 1:

<table>
<thead>
<tr>
<th>Dagens väder</th>
<th>Morgondagens väder</th>
</tr>
</thead>
<tbody>
<tr>
<td>☀</td>
<td>☀</td>
</tr>
<tr>
<td>☀</td>
<td>0.8</td>
</tr>
<tr>
<td>☀</td>
<td>0.2</td>
</tr>
<tr>
<td>☀</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Figur 4.2: Tabell över våra valda sannolikhet för morgondagens väder.

Dessa sannolikhet ska vi nu utnyttja för att ställa upp en tillståndmodell. I vårt fall finns tre möjliga tillstånd, \(S = \{\text{☀, ⛄, ☁}\}\), och varje dag finns en möjlighet att vi övergår från det aktuella tillståndet till ett annat (eller möjligen samma) enligt sannolikheterna i tabell 1. En tillståndsbeskrivning av detta återfinns i figur 4.3.
Innan vi går vidare tittar vi nu på ett exempel.

1. Givet att dagens väder är ☀, vad är sannolikheten för att morgondagens väder är ☂ och vädet i övermorgon ⛈?

Med sannolikhetera i tabell 1 och Markovs antagande i ekvation (2) får vi:

\[
P(q_2 = ☂, q_3 = ⛈ | q_1 = ☀) = P(q_3 = ⛈ | q_2 = ☂, q_1 = ☀) \cdot P(q_2 = ☂ | q_1 = ☀) = P(q_3 = ⛈ | q_2 = ☂) \cdot P(q_2 = ☂ | q_1 = ☀) = 0.05 \cdot 0.8 = 0.04
\]

Samma uträkning som ovan erhåller vi då vi går mellan tillstånden i figur 1 och multiplicerar med sannolikhetera långs vägen.

### 4.3 Dolda markovmodeller

Vi ska nu studera något som kallas *dolda markovmodeller*. Antag att vi har varit inlästa i ett rum i flera dagar och någon frågar oss hur vädet är. Den enda information vi har att tillgå är huruvida vakten som varje dag ger oss mat har med sig ett paraply (傘) eller ej (抜).

Låt oss anta sannolikhetera i nedanstående tabell: Sannolikheten för att vakten bär med sig ett paraply är 0.1 då det är soligt ute, 0.8 om det verkligen regnar och 0.3 om det är snöigt.

<table>
<thead>
<tr>
<th>Väder</th>
<th>Sannolikhet för paraply</th>
</tr>
</thead>
<tbody>
<tr>
<td>☀</td>
<td>0.1</td>
</tr>
<tr>
<td>⛈</td>
<td>0.8</td>
</tr>
<tr>
<td>☂</td>
<td>0.3</td>
</tr>
</tbody>
</table>

*Figur 4.4: Sannolikheten för att vakten bär med sig paraply beroende på väderförhållandet.*
Att beräkna sannolikheten för ett visst väder \( q_i \in \{\bigcirc, \bigstar, \blacklozenge\} \) måste nu helt baseras på observationen \( x_i \), där \( x_i = \bigcirc \) om vakten bär med sig ett paraply och \( x_i = \blacklozenge \) annars. Med hjälp av Bayes’ sats kan vi teckna denna villkorliga sannolikhet enligt följande:

\[
P(q_i | x_i) = \frac{P(x_i | q_i)P(q_i)}{P(x_i)}
\]  \( (4) \)

Om vi istället betraktar \( n \) dagar och en vädersekvens \( Q = \{q_1, q_2, \ldots, q_n\} \) samt en paraplysekvens \( X = \{x_1, x_2, \ldots, x_n\} \) får vi:

\[
P(q_1, \ldots, q_n | x_1, \ldots, x_n) = \frac{P(x_1, \ldots, x_n | q_1, \ldots, q_n)P(q_1, \ldots, q_n)}{P(x_1, \ldots, x_n)}
\]  \( (5) \)

Här känner vi igen \( P(q_1, \ldots, q_n) \) som sannolikheten för en markovkedja, i vårt fall en viss vädersekvens. På motsvarande sätt beskriver \( P(x_1, \ldots, x_n) \) sannolikheten (a priori) för att vi ska få se en viss paraplysekvens.

Sannolikheten \( P(x_1, \ldots, x_n | q_1, \ldots, q_n) \) kan estimeras med \( \prod_{i=1}^{n} P(x_i | q_i) \) såvida vi antar att \( x_i \) och \( q_i \), för alla \( i \), är oberoende av alla \( x_j \) och \( q_j \), där \( j \neq i \). Det innebär till exempel att huruvida vakten tar med sig ett paraply eller ej endast beror på dagens och inte tidigare (eller framtida) dagars väder.

Det vi vill åstadkomma är ett sätt att dra slutsatser om väder utomhus grundat på våra observationer (alltså huruvida vakten har med sig ett paraply eller ej). Därför kan vi bortse från a-priori-sannolikheten att se en viss paraplysekvens då den ju är oberoende av det väder vi vill förutsäga. Vi får då ett mått som är proportionellt mot den sanna sannolikheten. Detta mått kommer i fortsättningen betecknas \( L \) och kallas likelihoodvärde:

\[
P(q_1, \ldots, q_n | x_1, \ldots, x_n) \propto
\]

\[
L(q_1, \ldots, q_n | x_1, \ldots, x_n) = P(x_1, \ldots, x_n | q_1, \ldots, q_n) \cdot P(q_1, \ldots, q_n)
\]  \( (6) \)

Med Markovs förenkling (av första ordningen) får vi:

\[
P(q_1, \ldots, q_n | x_1, \ldots, x_n) \propto
\]

\[
L(q_1, \ldots, q_n | x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | q_i) \cdot \prod_{i=1}^{n} P(q_i | q_{i-1})
\]  \( (7) \)
4.3.1 Exempel

Låt oss nu studera ett par exempel där vi utnyttjar dolda markovmodeller för att göra utsagor om väder.


Först beräknar vi likelikelihoodvärdet för att den andra dagen är solig:

\[
L(q_2 = \emptyset | q_1 = \emptyset, x_2 = \uparrow) = P(x_2 = \uparrow | q_2 = \emptyset) \cdot P(q_2 = \emptyset, q_1 = \emptyset)
\]

\[
= 0.1 \cdot 0.8 = 0.08
\]

Samma sak för en regnig andra dag:

\[
L(q_2 = \bullet | q_1 = \emptyset, x_2 = \uparrow) = P(x_2 = \uparrow | q_2 = \bullet) \cdot P(q_2 = \bullet, q_1 = \emptyset)
\]

\[
= 0.8 \cdot 0.05 = 0.04
\]

Och slutligen för en snöig andra dag:

\[
L(q_2 = \ast | q_1 = \emptyset, x_2 = \uparrow) = P(x_2 = \uparrow | q_2 = \ast) \cdot P(q_2 = \ast, q_1 = \emptyset)
\]

\[
= 0.3 \cdot 0.15 = 0.045
\]

Med våra valda värden är alltså det mest troliga att även den andra dagen är solig (trots att vakten bär med sig ett paraply).

2. Antag nu att vi inte känner till väder när vi blir inlåsta. De följande tre dagarna kommer vakten utan paraply och vår uppgift är att beräkna likelihoodvärdet för att väder under dessa tre dagar har varit \{q_1 = \emptyset, q_2 = \ast, q_3 = \emptyset\}. Väder den dagen vi låstes in, q_0, känner vi alltså inte till och detta kommer att skapa problem då vi skall beräkna sannolikheten för att den första dagen är solig, d.v.s. \( P(q_1 = \emptyset | q_0) \). Vi måste därför bestämma oss för en initial sannolikhet och närmast till hands är att anta att alla väder är lika sannolika vilket i det här fallet ger oss:

\[
P(q_1 = \emptyset | q_0) = P(q_1 = \emptyset) = \frac{1}{3}.
\]

Likelihoodvärdet beräknas nu enligt följande:

\[
L(q_1 = \emptyset, q_2 = \ast, q_3 = \emptyset | x_1 = \ast, x_2 = \ast, x_3 = \ast) = \\
P(x_1 = \ast | q_1 = \emptyset) \cdot P(x_2 = \ast | q_2 = \ast) \cdot P(x_3 = \ast | q_3 = \emptyset) \cdot P(q_1 = \emptyset) \cdot P(q_2 = \ast | q_1 = \emptyset) \cdot P(q_3 = \emptyset | q_2 = \ast)
\]

\[
= 0.9 \cdot 0.7 \cdot 0.9 \cdot \frac{1}{3} = 0.0057
\]
4.3.2 Terminologi
Vi ska nu kortfattat beskriva en del beteckningar som ofta används när man talar om dolda Markovmodeller.

En dold markovmodell definieras av de möjliga tillståenden och en parametermängd \( \Theta = \{\pi, A, B\} \).

- \( S = \{s_1, s_2, \ldots, s_N\} \) beskriver de möjliga tillstånden (motsvarande de tre väderetyperna i texten ovan).
- \( V = \{v_1, v_2, \ldots, v_N\} \) beskriver de möjliga observationerna (motsvarande paraply eller ej paraply i texten ovan).
- \( \pi \) är en vektor som innehåller initialsannolikheter, där varje rad \( i \) innehåller sannolikheten för att \( s_i \) är det första tillståndet i en tillståndsekvens. I exempel 4.3.1 antog vi att dessa var lika stora, \( \pi_i = 1/N_s \) men man kan också tänka sig andra val.
- \( A \) är en matris som innehåller övergångssannolikheter mellan olika tillstånd (motsvarande figur 4.2 ovan)
- \( B \) är en matris som innehåller observationssannolikheter (Jfr figur 4.4). Där samlar vi alltså sannolikheten för att göra en viss observation (paraply eller ej paraply) när vi befinner oss i ett visst tillstånd (har ett visst väder).

När man använder sig av dolda Markovmodeller utgår man ifrån följande:
- Den (dolda) tillståndsekvensen \( Q = [q_1, q_2, \ldots, q_N] \), \( q_i \in S \).
- Observationssekvensen \( X = [x_1, x_2, \ldots, x_N] \), \( x_i \in V \).

4.3.3 Spaljédiagram
Ett så kallat spaljédiagram kan användas för att visualisera likelihoodberäkningar utifrån dolda Markovmodeller. Figur 4.5 visar ett sådant diagram för en modell med tre möjliga tillstånd.
Varje kolumn i spaljédiagrammet visar de möjliga tillstånden vid tidpunkten $n$. Samtliga tillstånd i en kolumn är ihopkopplade med tillstånden i den nästkommande kolumnen via övergångssannolikheterna i matris $A$ (endast synligt för tillstånd 1 vid tidpunkt 1 i figuren). Under spaljédiagrammet finns observationssekvensen $X = [x_1, x_2, ..., x_N]$ och $b_{i,k}$ beskriver sannolikheten för att man gör observationen $x_n = v_k$ i tillståndet $q_n = s_i$ vid tidpunkten $n$.

Låt oss nu återigen studera exemplet i 4.3.1. Ett spaljédiagram för vår dolda markovmodell som beskrev väderet syns i figur 4.6:
Likelihoodvärdet för en tillståndssekvens givet observationssekvensen finner vi nu enkelt genom att vandra genom spaljediagrammet och multiplicera samman övergångs- och observationssannolikheterna längs vår väg enligt följande:

\[
L = \pi_o \cdot b_{o,s_0} \cdot a_{s_0,s_1} \cdot b_{s_1,s_2} \cdot a_{s_2,s_3} \cdot b_{s_3,s_4} = 1/3 \cdot 0.9 \cdot 0.15 \cdot 0.7 \cdot 0.2 \cdot 0.9 = 0.0057
\]

4.4 Viterbiavkodning


4.4.1 Viterbialgoritmen

Viterbialgoritmen utnyttjar två variabler:

- \( \delta_n(i) \) är det högsta likelihoodvärdet som antas för någon väg bland de vägar som slutar i tillståndet \( i \) vid tidpunkten \( n \):
  \[
  \delta_n(i) = \max_{q_1,q_2,...,q_{n-1}} p(q_1,q_2,...,q_{n-1}, q_n = q_i, x_1,x_2,...,x_n \mid \Theta)
  \]  
  (8)

- \( \psi_n(i) \) är en variabel som används för att hålla reda på den bästa vägen som slutar i tillståndet \( i \) vid tidpunkten \( n \):
  \[
  \psi_n(i) = \arg\max_{q_1,q_2,...,q_{n-1}} p(q_1,q_2,...,q_{n-1}, q_n = q_i, x_1,x_2,...,x_n \mid \Theta)
  \]  
  (9)

Idén bakom viterbialgoritmen är att finna den mest sannolika vägen fram till alla mellanliggande och till sist det slutgiltiga tillståndet i spaljediagrammet. Vid varje tidpunkt \( n \) är det bara den mest sannolika vägen fram till tillstånd \( s_i \) som ”överlever”.

- **Initiering**
  \[
  \delta_1(i) = \pi_i \cdot b_{i,s_1} \quad i = 1,...,N_s
  \]
  \[
  \psi_1(i) = 0
  \]  
  (10)

där \( \pi_i \) betecknar en redan känd initialsannolikhet att befina sig i tillstånd \( s_i \) vid tiden \( n = 1 \).
• **Rekursion**

\[
\delta_n(j) = \max_{i \leq n} \{\delta_{n-1}(i) \cdot a_j \cdot b_j, x_j\}, \quad 2 \leq n \leq N
\]

\[
\psi_n(j) = \arg\max_{i \leq n} \{\delta_{n-1}(i) \cdot a_j \}, \quad 2 \leq n \leq N
\]

Optimalitet hittas genom att studera suboptimaliteter. Det gäller alltså att hitta den väg som ger det maximala likelihoodvärdet utifrån det högsta likelihoodvärdet i det närmast föregående steget samt övergångssannolikheterna till det nya tillståndet. Den bästa vägen hittas alltså via induktion.

• **Avbrott**

\[
p^*(X | \Theta) = \max_{i \leq N} \delta_N(i)
\]

\[
q_n^* = \arg\max_{i \leq N} \delta_N(i)
\]

Hitta det högsta likelihoodvärdet när sekvensen av observationer har nått sitt slut.

• **Backtracking**

\[
Q^* = [q_1^*, ..., q_N^*] \quad q_n^* = \psi_{n+1}(q_{n+1}^*) \quad n = N-1, N-2, ..., 1
\]

Avläs den bästa sekvensen av tillstånd från vektorerna \( \psi_n \).

### 4.4.2 Exempel

För att få grepp om hur viterbiavkodning fungerar ska vi avslutningsvis studera ett exempel. Som vanligt använder vi oss av vår vädermodell.

Antag att vi inte känner till väderet den dagen vi läses in. De tre första dagarna i fångenskap gör vi paraplyobservationerna: \( [\bigstar, \bigstar, \bigstar] \). Genom att utnyttja viterbialgoritmen vill vi nu finna den mest sannolika vädersekvensen givet denna observationssekvens. Som tidigare antar vi att alla väder är lika sannolika den första dagen.

1. **Initiering**

\[
\delta_1(\bigstar) = \pi_\bigstar \cdot b_{\bigstar, \bigstar} = 1/3 \cdot 0.9 = 0.3
\]

\[
\psi_1(\bigstar) = 0
\]

\[
\delta_1(\bigodot) = \pi_\bigodot \cdot b_{\bigodot, \bigodot} = 1/3 \cdot 0.2 = 0.067
\]
2. Rekursion

$n = 2$

Vi beräknar först likelihoodvärdet för att komma till tillstånd ⊙ från alla de tre möjliga föregående tillstånden och noterar den mest sannolika:

$$\delta_2(⊙) = \max(\delta_1(⊙) \cdot a_{⊙,\circ}, \delta_1(\bigstar) \cdot a_{\bigstar,\circ}, \delta_1(\bigtriangledown) \cdot a_{\bigtriangledown,\circ}) \cdot b_{⊙,\circ}$$

$$= \max(0.3 \cdot 0.8, \ 0.067 \cdot 0.2, \ 0.233 \cdot 0.2, \ 0.1 = 0.024$$

$$\psi_2(⊙) = ⊙$$

Likelihoodvärdet lagras i δ och det tillstånd som är den troligaste föregångaren i ψ. I ett spaljédiagram kan denna beräkning visualiseras enligt figur 4.7:

Figur 4.7: Beräkning av likelihoodvärde och troligaste föregångare då vi hamnar i tillstånd ⊙ vid tidpunkt 2.
Vi utför nu analoga beräkningar för tillstånden ◆ och ✿:

\[
\delta_2(◆) = \max(\delta_1(⊙) \cdot a_{⊙, ◆}, \delta_1(◆) \cdot a_{◆, ◆}, \delta_1(✻) \cdot a_{✻, ◆}) \cdot b_{◆, ✿} = \max(0.3 \cdot 0.05, 0.067 \cdot 0.6, 0.233 \cdot 0.3) \cdot 0.8 = 0.056
\]

\[
\varphi_2(◆) = ✿
\]

\[
\delta_2(✽) = \max(\delta_1(⊙) \cdot a_{⊙, ✿}, \delta_1(◆) \cdot a_{◆, ✿}, \delta_1(✻) \cdot a_{✻, ✿}) \cdot b_{◆, ✿} = \max(0.3 \cdot 0.15, 0.067 \cdot 0.2, 0.233 \cdot 0.5) \cdot 0.3 = 0.035
\]

\[
\varphi_2(✽) = ✿
\]

\[n = 3\]

\[
\delta_3(⊙) = \max(\delta_2(⊙) \cdot a_{⊙, ⊙}, \delta_2(◆) \cdot a_{◆, ⊙}, \delta_2(✻) \cdot a_{✻, ⊙}) \cdot b_{⊙, ✿} = \max(0.024 \cdot 0.8, 0.056 \cdot 0.2, 0.035 \cdot 0.2) \cdot 0.1 = 0.0019
\]

\[
\varphi_3(⊙) = ⊙
\]

\[
\delta_3(◆) = \max(\delta_2(⊙) \cdot a_{⊙, ◆}, \delta_2(◆) \cdot a_{◆, ◆}, \delta_2(✻) \cdot a_{✻, ◆}) \cdot b_{◆, ✿} = \max(0.024 \cdot 0.05, 0.056 \cdot 0.6, 0.035 \cdot 0.3) \cdot 0.8 = 0.0269
\]

\[
\varphi_3(◆) = ◆
\]

\[
\delta_3(✽) = \max(\delta_2(⊙) \cdot a_{⊙, ✿}, \delta_2(◆) \cdot a_{◆, ✿}, \delta_2(✻) \cdot a_{✻, ✿}) \cdot b_{◆, ✿} = \max(0.024 \cdot 0.15, 0.056 \cdot 0.2, 0.035 \cdot 0.5) \cdot 0.3 = 0.0052
\]

\[
\varphi_3(✽) = ✿
\]

\[
\delta_3(⊙) = 0.0019
\]

\[
\delta_3(◆) = 0.0269
\]

\[
\delta_3(✽) = 0.0052
\]

Figur 4.8: Spaljédiagram där den mest sannolika vägen fram till varje sluttillstånd är markerad.
3. Avbrott

För att hitta den globalt mest troliga vägen genom spaljédiagrammet måste vi först finna det sista tillståndet i den mest troliga sekvensen:

\[ P^* (X | \Theta) = \max (\delta_i (i)) = \delta_3 (\bullet) = 0.0269 \]
\[ q_3^* = \arg \max (\delta_3 (i)) = \bullet \]

4. Backtracking

Den bästa sekvensen kan nu hittas i \( \psi \)-vektorerna. (Se figur 4.9)

\[ n = N - 1 = 2 \]
\[ q_2^* = \psi_2 (q_3^*) = \psi_2 (\bullet) = \bullet \]

\[ n = N - 2 = 1 \]
\[ q_1^* = \psi_1 (q_2^*) = \psi_2 (\bullet) = \diamond \]

\[ \delta_3 (\circ) = 0.0019 \]
\[ \delta_3 (\bullet) = 0.0269 \]
\[ \delta_3 (\diamond) = 0.0052 \]

**SEKVENS:**

\[ x_1 = \circ \]
\[ x_2 = \diamond \]
\[ x_3 = \diamond \]

\[ n=1 \quad n=2 \quad n=3 \]

**Figur 4.9: Spaljédiagram som visar den mest sannolika vädersekvensen.**

Den mest sannolika vädersekvensen är alltså:

\[ Q^* = [q_1^*, q_2^*, q_3^*] = [\diamond, \bullet, \bullet] \]
5 Det nya systemet


5.1 Förutsättningar

Tanken är att vi ska finna en avkodningsstrategi som drar fördel av den statistik vi samlade in i kapitel 3. Inledningsvis kommer vi att utnyttja kunskapen om markovmodeller som vi skaffade oss i föregående kapitel. Lite längre fram ska vi se att vårt tidigare påstående att textinmatnings-problemet lämpar sig mycket väl för att modelleras med hjälp av dolda markovmodeller, är riktigt. Vi kommer genomgående att använda oss av samma beteckningar som tidigare. Som vanligt förutsätter vi att all inmatning sker från en traditionell knappsats som den i figur 2.1.

5.1.2 Utförande

Samtliga ansatser som beskrivs i det här kapitlet har implementerats och utvärderats i matlab.

5.1.3 Testtext

För att utvärdera de olika systemansatser vi kommer att göra ska vi utnyttja en testtext bestående av 126 tecken inklusive blanksteg. Denna testtext är bara ett redskap för att vi ska kunna göra enkla jämförelser mellan de olika tillvägagångssätten som undersöks inledningsvis. I senare kapitel kommer vi att utnyttja ett mer omfattande testmaterial för våra undersökningar. Den testtext vi ska använda oss av i den inledande undersökningen är ett fritt översatt citat av Albert Einstein:

**JAG VET EJ MED VILKA VAPEN DET TREDJE VÄRLDSKRIGET KOMMER ATT UTKÄMPAS MEN DET FJÄRDE KOMMER ATT UTKÄMPAS MED PÅKAR OCH STENAR**

5.1.4 Felmått

Eftersom den inledande undersökningen endast involverar jämförelser mellan olika ansatser krävs inget sofistikerat felmått. Måttet på hur ”bra” ett system är kommer tills vidare bara bero på det antal symboler som blir korrekta då testtexten används som indata.
5.1.5 Begränsningar

I målet för detta examensarbete ingår inte att studera användargränssnitt. Alla tester och jämförelser mellan olika system kommer därför att ske på en grundnivå där ingen interaktion med användaren sker utöver de rena knapptryckningarna. Det ges alltså ingen möjlighet till utomstående korrigering av eventuella fel.

5.2 Textinmatningsproblemet som dold markovmodell

Vi påminner om att dolda Markovmodeller kan utnyttjas för att göra utsagor om troliga tillståndssekvenser utifrån en viss observationssekvens. När det gäller vårt textinmatningssystem har vi i det enklaste fallet 30 tillstånd, ett för varje symbol:

\[ S = \{A, B, C, \ldots, Ò, ø, -\} \]

Den enda information systemet kommer att ha tillgång till för att avgöra vilken symbol användaren vill mata in, är vilken knapp som trycktes. Våra observationssekvenser kommer alltså att bestå av element från mängden med knappar, i vårt fall nio stycken:

\[ X = [x_1, x_2, \ldots, x_n], x_j \in V \text{ där} \]

\[ V = \{2, 3, 4, 5, 6, 7, 8, 9, 0\} \]

B-matrisen med observationssannolikheter kommer i vårt fall att se lite speciell ut. Givet ett visst tillstånd (d.v.s. en viss symbol) råder det nämligen ingen tvekan om vilken observation som gjordes. Hamnar vi till exempel i tillståndet T vet vi att knapp 8 trycktes ned (såvida mobiltelefonen inte är defekt). Vi vet också att ingen annan knapp kan ha trycks ned och får därför bara ettor och nollor i B-matrisen.


5.2.1 Beteckningar

5.3 Kontextmodeller

Vi har nu alla verktyg vi behöver för att inleda vår jakt på ett sannolikhetsbaserat textinmatningssystem. Avsikten är att vi ska börja med ett väldigt enkelt system för att sedan utföra successiva förbättringar tills vi når tillfredsställande testdata. Vår första ansats kommer att bygga på den mest primitiva kontextmodellen, nämligen den av ordning 0.

5.3.1 Ordning 0

I en kontextmodell av ordning 0 tar vi ingen som helst hänsyn till tidigare inmatning (vilket tillstånd vi kommer ifrån). Istället är det bara den direkta sannolikheten att hamna i ett visst tillstånd som styr. I vårt textinmatningssystem får detta konsekvensen att rena bokstavsfrekvenser (se Appendix A.1) får avgöra vilket gissning vi gör. Vid varje knapptryckning väljs den mest sannolika bland knappens bokstäver som utdata.

Om vi matar in vår testtext till ett system som bygger på en kontextmodell av ordning 0 får vi följande utdata:

LAI TET EL NEE TILLA TAREN EET TREELE TARLERLRIET LNNNER ATT TTLANRAR NEN EET ELAREE LNNNER ATT TTLANRAR NEE RALAR NAI RTENAR

Antalet korrekta symboler är 72 och antalet felaktiga således 54. Precis som väntat blir resultatet inte särskilt bra och vi inser att en modell som inte tar hänsyn till tidigare inmatning är dömd att misslyckas. Vi utvecklar därför vårt system en aning och studerar en kontextmodell av ordning 1.

5.3.2 Ordning 1


Det som återstår att bestämma innan vi kan testa vår modell är värdena i A-matrisen, d.v.s. övergångssannolikheterna mellan olika tillstånd. För att bestämma dessa skall vi utnyttja den statistik över bigram i svenska språket som vi samlat in tidigare. Sannolikheten för att gå ifrån tillstånd \( q_i \) till \( q_j \) motsvaras helt enkelt av den \textit{relativa frekvensen} för bigrammet som binder
ihop tillstånden. Relativ frekvens syftar i det här fallet på frekvensen för bigrammet relativt alla andra bigram som inleds med samma bokstav. Om vi till exempel står i tillståndet A kommer sannolikheten att hamna i tillståndet T att vara den relativa frekvensen för bigrammet AT.

Med vår testtext som indata erhåller vi nu följande:

LAG VET DL MED VILLA VAREN DET VSEDLE VARKERKRIGET KOMMER ATT VULANSAR MEN DET DLARED KOMMER ATT VULANSAR MED SALAR MAG STENAR

Vi ser en klar förbättring jämfört med den tidigare modellen och får nu 89 korrekta och 37 felaktiga symboler. Trenden verkar lovande och vi ökar modellens ordning ytterligare.

5.3.3 Ordning 2


En förändring av tillstånden leder givetvis till att vi även måste förändra A- och B-matrismatriserna.

Övergångssannolikheterna finner vi som den relativa frekvensen hos det trigram som binder ihop två tillstånd. Sannolikheten för att gå ifrån till exempel tillstånd HE till EJ motsvaras alltså av den relativa frekvensen för trigrammet HEJ. Denna strategi innebär också vi kommer få en hel del nollor i A-matrissen. För att en övergång mellan tillstånd $q_i$ till $q_j$ skall vara möjlig måste ju $q_i$ avslutas med den symbol som inleder $q_j$. Om inte detta är fallet blir övergångssannolikheten 0. Dessutom blir sannolikheten givetvis 0 i de fall då två tillstånd binds ihop av ett trigram som inte förekommer någon gång i vår statistik.

Som vi tidigare konstaterat har B-matrisen med observationssannolikheter ettor vid de tillstånd som är möjliga, givet en viss observation. Här innebär det konkret alla de tillstånd som avslutas med någon av symbolerna från den observerade knapptryckningen.

Då vi återigen använder vår testtext som indata får vi följande:
LAG VET DJ OFF VILLA VAREN DET VREDLE VARKERKRIGET KOMMER ATT VULANSAR MEN DET DJARED KOMMER ATT VULANSAR MED SÅLAR MAG STENAR

Förbättringen jämfört med vår tidigare modell är minimal och vi får 90 korrepta och 36 felaktiga symboler. Vad beror detta på? Ja, förmodligen kan vi inte komma särskilt mycket längre med kontextmodellering när vi betraktar ren inmatning utan interaktion från en användare. Förutsägelserna bygger på det bigram som föregår inmatningen och det är givetvis avgörande om detta bigram är korrekt eller ej. Utan korrigering från användaren riskerar vi att få fel som fortpflanter sig framåt i texten.

Att en kontextmodell av ordning 2 trots allt fungerar relativt väl när korrigering tillåts är Letterwise ett lysande exempel på. Då man med säkerhet bygger förutsägelserna på korrekt information visar det sig att systemet gissar rätt symbol i en hög procent av fallen.[9]

Eftersom vår undersökning bara berör ren inmatning tvingas vi dock inse att ytterligare kontextmodellering inte kan förbättra vårt system närmast, även om vi höjer modellordningen. (I själva verket blev testdata sämre för en kontextmodell av ordning 3, med 76 korrepta och 50 felaktiga symboler). Istället skall vi nu försöka finna den optimala tillståndssekvensen med hjälp av viterbiavkodning.

5.4 Viterbiavkodning

Givet en observationssekvens (knapptryckningssekvens) finner alltså viterbialgoritmen den mest sannolika tillståndssekvensen (symbolsekvensen). Vi kommer direkt att applicera viterbiavkodning på den kontextmodell av ordning 2 som vi just tog fram.

5.4.1 Viterbiavkodning för en kontextmodell av ordning 2

Vi använder nu viterbiavkodning på den knapptryckningssekvensen som motsvarar vår testtext. I en kontextmodell av ordning 2 får vi då följande optimala symbolsekvens:

**JAG VET FL MED TILLA VAREN DET UREDJE VÄRLDSKRIGET KOMMER ATT UTLÄMPAR MEN DET FLARED KOMMER ATT UTLÄMPAR MED PÄKAR OCH STENAR**


### 5.4.2 Viterbiavkodning för en kontextmodell av ordning 3

Precis som tidigare kommer vi att utöka tillståndersrepresentationen i vår modell så att den motsvarar en tredje ordningens kontextmodell. Utvidgningen blir helt analog med det vi gjort tidigare. I vår nya modell får varje **trigram** motsvara ett eget tillstånd vilket totalt ger oss 27000 tillstånd. Övergångar mellan olika tillstånd kan bara ske när de två symboler som avslutar det tillstånd man kommer ifrån, inleder det tillstånd man är på väg till. Sannolikheterna för dessa övergångar bestäms via relativ frekvenser för tetragram. Till exempel avgörs alltså övergångsannolikheten från tillstånd **ANK** till **NKA**, av den relativa frekvensen för tetragrammet **ANKA**.

Initialsannolikheterna ställer som vanligt till problem. Vi vill återigen ge systemet information om att varje inmatning inleds med ett nytt ord. På något vis skall alltså den relativa frekvensen för bigram som inleds med blanksteg och avslutas med den första knapptryckningens symboler, avgöra initialsannolikheterna. Låt oss titta på ett exempel för att åskådliggöra detta:

Låt oss anta att den första knapp som användaren trycker på är knapp 5 med bokstäverna J, K och L. Initialsannolikheterna för de olika starttillstånden kommer då vara lika med den relativa frekvensen för bigrammen _J, _K och _L. Ett problem kvarstår dock. Tillstånden vi betraktar nu består ju av trigram och vi måste fortfarande definiera vilka de möjliga starttillstånd som våra initialsannolikheter hänvisar till, år. I ett allmänt fall låter vi därför alla trigramtillstånd som avslutas med _* vara möjliga starttillstånd, där * motsvarar den första knapptryckningens symboler. I det här fallet blir alltså de möjliga starttillstånden de 90 trigram som avslutas med _J, _K eller _L.

Resultatet vid inmatning av testtexten blir följande:

**JAG VET EJ MED TILLA VAREN DET VREDJE VÄRLDSKRIGET KOMMER ATT UTLÄMPAR MEN DET FJÄRDE KOMMER ATT UTLÄMPAR MED RÅKAR OCH STENAR**

117 korrekta och endast 9 felaktiga symboler gör att vi med lite fantasi nu nästan kan utläsa hela meddelandet. Precis som vi hoppades fick vi en klar förbättring genom att utöka kontextmodellens ordning.
5.5 Val av system


Att vi ändå inte genomför någon ytterligare utökning av vår modell beror främst på att vi inte har skaffat oss tillgång till statistik över pentagram-frekvenser som skulle krävas för att generera övergångssannolikheter för en modell av ordning 4. Dessutom betraktar vi våra testresultat från tredje ordningens kontextmodell som tillräckligt tillfredsstillande.

Den strategi vi kommer att utnyttja i återstoden av rapporten kommer därför att vara viterbiavkodning för en kontextmodell av ordning 3.

5.5.1 Beteckning

För att underlätta den fortsatta texten kommer systemet härdanefter betecknas Qtap.

5.5.2 Utvärdering

En stor fördel hos viterbiavkodning jämfört med vanlig kontextmodellering är att systemet inte gissar utdata förrän inmatningen är avslutad. Varje symbolgissning kan på så sätt baseras på information om både tidigare och efterkommande inmatningar. I kontextmodellerna däremot, gissar vi symbolerna efter hand och kan således inte ta hänsyn till efterkommande inmatningar. Detta är en betydande begränsning eftersom ett felaktigt val aldrig kan korrigeras av systemet självt. Det är av denna anledning Letterwise är så beroende av att användaren hela tiden ser till att systemets utdata motsvarar önskad inmatning. För att understycka viterbialgoritmens fördel gentemot vanlig kontextmodellering studerar vi ett kort exempel:

Låt oss anta att vi vill mata in ordet DAG med en traditionell knappsats. Knappsekvensen för DAG är då ØØ Ø. Nedanstående tabell visar hur utdata

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Sannolighet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ø</td>
<td>0.1</td>
</tr>
<tr>
<td>ØØ</td>
<td>0.2</td>
</tr>
<tr>
<td>ØØØ</td>
<td>0.3</td>
</tr>
</tbody>
</table>

För att viterbiavkodning ska fungera korrekt måste systemet kunna återvända till ett ensamt tillstånd från vilket det kommer att utföra en sannolikhetsexamen. I detta fall skulle vi välja tillståndet ØØ Ø som det mest sannolika av alla möjliga tillstånd. Denna strategi kan dock inte alltid fungera, och det är därför viktigt att medfölja med en förlorad utdata som vi inte kan återvända till ett ensamt tillstånd.
Förenklad textinmatning på mobila enheter med hjälp av kontextbaserad språktolkning

blir för en kontextmodell av ordning 2, beroende på om vi utnyttjar viterbiavkodning eller ej.

<table>
<thead>
<tr>
<th>Inmatning</th>
<th>Ren kontextmodell</th>
<th>Viterbiavkodning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>EC</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>ECI</td>
<td>DAG</td>
</tr>
</tbody>
</table>

Figur 5.1: Utdata vid inmatning av ordet ”DAG” beroende på om viterbiavkodning används eller ej.

Vi ser tydligt hur den rena kontextmodellen inte kan ångra tidigare val och därför tvingas välja den ovanliga bokstavskombinationen ECI som utdata trots att knappsekvensen kan motsvara DAG. Med viterbiavkodning får vi inte detta bekymmer eftersom den optimala tillståndssekvensen väljs när inmatningen är avslutad.


För våra ändamål, där vi endast ämnar betrakta systemet på en nivå som inte involverar användarinteraktion, betraktar vi dock viterbiavkodning för en kontextmodell av ordning 3 som en tillfredsställande strategi. Vårt testresultat med 117 korrepta symboler bedömer vi vara tillräckligt bra och är nu redo för en undersökning huruvida en förändring av knappsatsens utseende kan förbättra systemet ytterligare.
6 Förändring av knappsatsen


6.1 Bakgrund

I målet för vår undersökning ingick inte bara att finna en lämplig strategi för ett sannolikhetbaserat textinmatningssystem utan också att undersöka vad en förändring av symbolernas placering på knappsatsen har för inverkan på systemets prestanda. Utseendet på dagens knappsatser härrör från en tid då enkelheten med bokstäverna i bokstavsordning var mer tilltalande än hög effektivitet vid inmatning. Då behovet av effektiv textinmatning ökat så markant de senaste åren börjar denna uppfattning förändras.

6.2 Förutsättningar

6.2.1 Inmatningssystem

Genom hela kapitlet kommer vi att förutsätta att textinmatningssystemet som används är det vi tog fram i föregående kapitel, dvs Qtap.

6.2.2 Testdata

I sökandet efter bra knappsatser kan vi inte längre utnyttja endast vår korta testtext från kapitel 5 eftersom vi då riskerar att finna en knappsats som fungerar väldigt bra för just den texten, men inte nödvändigtvis i ett allmänt fall. Därför kommer vi att använda flera olika testtexter som vi anpassar efter våra ändamål. Dessa kommer att betecknas med alfabetets bokstäver:

<table>
<thead>
<tr>
<th>Testtext</th>
<th>Antal tecken</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>126 tecken</td>
</tr>
<tr>
<td>B</td>
<td>196 tecken</td>
</tr>
<tr>
<td>C</td>
<td>1558 tecken</td>
</tr>
<tr>
<td>D</td>
<td>529 tecken</td>
</tr>
<tr>
<td>E</td>
<td>100 tecken</td>
</tr>
<tr>
<td>F</td>
<td>1000 tecken</td>
</tr>
<tr>
<td>G</td>
<td>10.000 tecken</td>
</tr>
<tr>
<td>H</td>
<td>100.000 tecken</td>
</tr>
</tbody>
</table>

Figur 6.1: Antal tecken för de olika testtexter som utnyttjas i kapitlet.

6.2.3 Felmått

Precis som tidigare kommer vi att låta antalet symbolfel avgöra hur ”bra” en knappsats är.

6.3 Dagens knappsats

På den traditionella knappsatsen som vi hittills utgått ifrån, är bokstäverna placerade i bokstavsordning över knapparna och vi kan lätt skaffa oss en känsla av att denna placering inte är optimal ur inmatningssynpunkt. Betrakta till exempel knapp 9 där bokstäverna WXYZ återfinns. Samtliga dessa bokstäver är relativt ovanliga i det svenska språket och knappen används därför långt mer sällan än andra knappar vid normal inmatning. Dessutom är till exempel R och S väldigt vanliga bokstäver i det svenska språket och möjligen är det olyckligt att ha dessa placerade på samma knapp. De sammanlagda sannolikheterna för varje knapp syns i figur 6.2 och visar att fördelningen troligen kunde vara bättre:

![Sannolikhetsfördelning mellan knapparna](image-url)

Figur 6.2: Sannolikhetsfördelning mellan knapparna på en traditionell knappsats.
En mer genomtänkt placering av bokstäverna skulle kunna innebära förbättringar för både ordlistebaserade och sannolikhetsbaserade system. Hos t9 skulle till exempel en sådan förändring kunna minska antalet kollisioner, dvs ord som skrivs med samma knappsekvens.

### 6.3.1 Test

Innan vi inleder vårt sökande efter en lämplig knappsats ska vi undersöka vilka testdata vi får när vi utnyttjar den traditionella knappsatsen. Med våra fem första testtexter som indata får vi följande resultat:

<table>
<thead>
<tr>
<th></th>
<th>Fel</th>
<th>Rätt</th>
<th>Felprocent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testtext A</td>
<td>9</td>
<td>117</td>
<td>7.1</td>
</tr>
<tr>
<td>Testtext B</td>
<td>24</td>
<td>172</td>
<td>12.2</td>
</tr>
<tr>
<td>Testtext C</td>
<td>79</td>
<td>1479</td>
<td>5.1</td>
</tr>
<tr>
<td>Testtext D</td>
<td>61</td>
<td>468</td>
<td>11.5</td>
</tr>
<tr>
<td>Testtext E</td>
<td>5</td>
<td>95</td>
<td>5.0</td>
</tr>
<tr>
<td>Totalt</td>
<td>178</td>
<td>2509</td>
<td>7.1</td>
</tr>
</tbody>
</table>

Figur 6.3: Testresultat för den traditionella knappsatsen.

Hur Testtext A ser ut då vi använder den traditionella knappsatsen har vi redan sett i avsnitt 5.4.2 och avstår ifrån en upprepning.

### 6.4 Intuitiv symbolplacering

Vi har konstaterat att dagens knappsats förmodligen går att förbättra men frågan är hur en sådan förbättring ska gå till. Låt oss för ett ögonblick se vart ett enkelt resonemang kan leda oss. Vi anar från vår tidigare diskussion att sannolikhetsfördelningen över de olika knapparna förmodligen bör vara jämnare än vad som är fallet med den traditionella knappsatsen. Dessutom vill vi se till att inga av de vanligaste bokstäverna i det svenska språket hamnar på samma knapp, som är fallet med till exempel R och S på den traditionella knappsatsen. För att åstadkomma detta utgår vi ifrån rena bokstavsfrekvenser (se Appendix A.1) och placera först de nio vanligaste symbolerna på varsin knapp. De övriga symbolerna placera vi sedan ut så att sannolikhetsfördelningen blir relativt jämn. En sådan knappsats skulle till exempel kunna se ut enligt figur 6.4:
Förenklad textinmatning på mobila enheter med hjälp av kontextbaserad språktolkning

Figur 6.4: Exempel på knappsats där vårt intuitiva resonemang ligger till grund för placeringen av symboler.

Med denna nya knappsats får de sammanlagda sannolikheterna följande utseende:

![Sannolikhetsfördelning mellan knappar](chart.png)

Figur 6.5: Sannolikhetsfördelning mellan knapparna på vår intuitiva knappsats.

Vi får en nödvändig avvikelse för den sista knappen eftersom blanksteg är så vanligt men de övriga knapparna har nu ett mer tilltalande utseende än tidigare.
6.4.1 Test
Vi matar nu in våra fem testtexter med denna knappsats och får följande data:

<table>
<thead>
<tr>
<th></th>
<th>Fel</th>
<th>Rätt</th>
<th>Felprocent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testtext A</td>
<td>9</td>
<td>117</td>
<td>7.1</td>
</tr>
<tr>
<td>Testtext B</td>
<td>8</td>
<td>188</td>
<td>4.1</td>
</tr>
<tr>
<td>Testtext C</td>
<td>33</td>
<td>1525</td>
<td>2.1</td>
</tr>
<tr>
<td>Testtext D</td>
<td>47</td>
<td>482</td>
<td>8.8</td>
</tr>
<tr>
<td>Testtext E</td>
<td>7</td>
<td>93</td>
<td>7.0</td>
</tr>
<tr>
<td>Totalt</td>
<td>104</td>
<td>2509</td>
<td>4.1</td>
</tr>
</tbody>
</table>

Figur 6.6: Testresultat för den intuitiva knappsatsen.

Vi kan konstatera att testtext A återigen innehöll 9 felaktiga symboler:

**JAG VET EJ MED VILKA VAREN DET TREDJE VÄRLDSKRISET KOMMER ATT UTKÄRRAS MEN DET FLÄMDE KOMMER ATT UTKÄRRAS MED RÅKAR OCH STENAR**

Däremot kan vi se att vår intuitiva knappsats gav en förbättring om vi betraktar alla fem testtexter. Slutsatsen vi drar av detta är främst att sökandet efter alternativa knappsatser är motiverat och vi ska nu försöka finna en mer organiserad metod för att hitta lämpliga symbolplaceringar.

6.5 Systematisk symbolplacering
Frågan vi nu ska ställa oss är alltså om vi kan hitta något systematiskt sätt att fördela symbolerna över knapparna. De önskemål vi hade om jämn sannolikhetsfördelning och spridning av vanliga bokstäver kan vara intuitivt tilltalande men vi har inte några direkta belägg för att detta är nödvändiga egenskaper hos en bra knappsats. Därför måste vi nu närmare undersöka vad vi bör ta hänsyn till när vi väljer placering av symbolerna.

Det system vi utarbetade i föregående kapitel finner den optimala sekvensen av symboler givet knapptryckningarna. Det innebär att både tidigare och framtida inmatningar kommer att ha betydelse för systemets val av utdata. Givet en viss omgivning och en viss knapptryckning vill vi alltså att systemet så enkelt som möjligt skall kunna välja utdata. Vi inser att det inte nödvändigtvis är hur vanliga knappens symboler är som är den avgörande faktorn, utan snarare om de har liknande funktion. Bokstäver med liknande funktion syftar i det här fallet på bokstäver som förekommer i liknande omgivningar. Låt oss studera ett kort exempel:

Många substantiv i bestämd form i svenska slutar på **N** eller **T**. Dessa bokstäver kommer därför ofta förekomma i liknande omgivningar och det


6.5.1 Beräkning av språkliga funktioner

Vi definierade tidigare ”liknande funktion” som ”förekommande i liknande omgivningar”. För att få ett mått på en bokstavs funktion bör vi alltså studera vilka omgivningar den förekommer i. Till detta ändamål utnyttjar vi vår statistik över trigram som vi samlade in i kapitel 3. Vi kommer att läsa mittensymbolen i trigrammen till den bokstav vi vill studera och låter de andra två symbolerna utgöra ”omgivningen”. Detta ger oss 900 möjliga omgivningar för varje symbol och vi kommer därför att införa ett 900-dimensionellt rum att utföra våra beräkningar i. Vi inför också skalärprodukt och 2-vektornorm för att kunna göra euklidska beräkningar av ”avstånd” och ”mellanliggande vinkel”. Varje symbols position i rummet kommer att definieras av en vektor som innehåller antalet förekomsterna i PAROLE för de trigram som har den aktuella symbolen i mitten.

Vi ska titta på ett enkelt exempel för att visa hur detta fungerar principiellt. För åskärligheten skull studera vi ett alfabet som endast består av de tre symbolerna A, B och C. Istället för 900 får vi då $3^2 = 9$ dimensioner. För att bestämma varje bokstavs position i rummet skall vi alltså studera antalet förekomster hos de trigram som har respektive bokstav i mitten.
Vektorerna som anger bokstävernas position i rummet motsvaras direkt av de högra kolumnerna i figur 6.7. Resultatet verkar intuitivt rimligt. Vi ser till exempel att A främst förekommer omgiven av konsonanter i vårt reducerade alfabet.

Vi återgår nu till vårt 900-dimensionella fall och försöker finna ett bra sätt att jämföra språkliga funktionen mellan olika bokstäver.

### 6.5.2 Avstånd


Vi inser att avståndet inte är något bra mått när vi vill jämföra bokstävers olika funktion. Snarare är det vilka riktningar våra vektorer har i rummet som är intressant och därför ska vi fortsättningsvis studera vinklar.
6.5.3 Vinkelavstånd

Med hjälp av vår skalärprodukt kan vi enkelt bestämma vinkeln mellan två vektorer och de inbördes vinklarna mellan samtliga symboler finns i appendix B. Där finns också en tabell som för varje symbol visar vilka andra symboler som enligt vår metod används på liknande sätt. De misstankar vi hade tidigare bekräftats av tabellen. För alla vokaler gäller till exempel att de andra vokalerna återfinns bland de mest likartade symbolerna. Vi kan också se att N och T förekommer i liknande omgivningar, precis som vi trodde.

Vi har skaffat oss vad vi tror är en ganska bra bild av vilka symboler som används på liknande sätt i det svenska språket. Frågan är hur vi skall utnyttja detta i vår symbolplacering. Önskvärt vore troligen att utföra någon slags optimering med hjälp av dator men innan vi funderar på det bör vi undersöka om vinklarna mellan våra vektorer verkligen ger oss ett lämpligt fördelningskriterium i jakten på bra knappsatser.

6.5.4 Test

Vi ska utföra en testserie som omfattar 100 försök och utnyttjar testtext \( F \), som alltså består av 1000 tecken. Inför varje försök låter vi en dator slumpa fram den symbolplacering som skall användas för inmatning och noterar antalet felaktiga symboler som knappsatsen ger. Vi beräknar också medelvinkelavståndet mellan symbolerna på varje knapp och summerar dessa. Utifrån detta bildar vi sedan det genomsnittliga vinkelavståndet för hela knappsatsen genom att dividera summan med antalet knappar. Vår förhoppning är att bra knappsatser som ger få felaktiga symboler skall ha ett högt genomsnittligt vinkelavstånd. Testserien utföll enligt figur 6.8:

Figur 6.8: Genomsnittligt vinkelavstånd och felfrekvens för 100 framslumpade knappsatser.
6.5.5 Utvärdering

Resultatet från vår testserie är inte särskilt upplyftande. Vi kan möjligen ana en mycket svagt avtagande medeltrend men lutningen är inte lika tydlig som vi hade hoppats på. Till exempel kan vi konstatera att knappsatser _kan_ ge mycket bra resultat trots att vinkelavståndet är förhållandevis litet. Den allra bästa knappsatsen från försöken hade till exempel ett vinkelavstånd (1,1413) som var lägre än genomsnittet.

Vad kan vi då dra för slutsatser av vår testserie? Ja, klart är att vårt fördelningskriterium med vinkelavstånd inte var tillräckligt för att enkelt generera bra knappsatser. Den främsta anledningen till detta är förmodligen att vinkelavståndet inte ensamt duger som kriterium för att finna bra knappsatser. Att symbolplaceringar med relativt lågt värde på det genomsnittliga vinkelavståndet fortfarande kan ge väldigt bra testresultat antyder att det finns fler faktorer som avgör. Vi kan emellertid också konstatera att det bland de knappsatser som hade stort genomsnittligt vinkelavstånd, inte finns någon som gav riktigt usla testresultat. Vi har därför stora förhoppningar på att vårt mått på språklig funktion ger spännande möjligheter och inbjuder till vidare studier. Mer kommentarer kring detta finns i avsnitt 8.2.4.

Vi avslutar nu försöken att systematiskt placera symbolerna och övergår till att använda ren datorkraft i vår jakt på bra knappsatser. Vi kommer dock att ha anledning att återkomma till vinkelavstånd när vi skall utvärdera den knappsats vi finner den vägen.

6.6 Datorstödd symbolplacering

Då vårt försök till systematisk placering inte utmynnade i särskilt goda testresultat skall vi nu försöka finna en bra knappsats med datorns hjälp. Vi kommer att slumpa fram ett stort antal knappsatser i matlab och precis som på den traditionella knappsatsens begränsar vi oss till nio knappar.

Eftersom vi vill reservera testtext A-E till att vara kontrolltexter skall vi inte utnyttja dessa i våra datorstödda försökerier.

6.6.1 Försöksserie 1


Det principiella utseendet hos den fördelning vi erhåller visas i figur 6.9, där den traditionella knappsatsens resultat (40 fel) har markerats med en prick.
Förenklad textinmatning på mobila enheter med hjälp av kontextbaserad språktolkning

Försök 1

Figur 6.9: Fördelningen av knappsatser i försöksserie 1, med avseende på antalet felaktiga symboler. Den traditionella knappsatsens testresultat har markerats med en prick.

Vi kan konstatera att den traditionella knappsatsen ger något bättre resultat än det genomsnittliga resultatet för våra framslumpade knappsatser. Trots detta finns det ändå ett ganska stort antal som ger färre felaktiga symboler. Den bästa knappsatsen i denna försöksserie gav endast 8 fel medan motsvarande siffra för den sämsta knappsatsen var 126.

6.6.2 Försöksserie 2

Vi gör nu en första gallring och väljer ut de 3000 bästa knappsatserna från försöksserie 1. Dessa motsvaras i figur 6.9 av knappsatser med färre än 36 felaktiga symboler.
I vår nästa försöksserie skall vi använda en längre testtext för att försäkra oss om att knappsatserna verkligen är bra i en vid mening och inte bara lämpliga för vår relativt korta testtext i försöksserie 1. Testtexten vi använder nu är testtext G som består av 10.000 tecken. Detta bör ge oss en god fingervisning om vilka knappsatser vi vill studera noggrannare.

Resultatet för de 3000 knappsatser vi undersökte syns i figur 6.11. Även här har vi matat in testtexten med den traditionella knappsatsen och markerat resultatet (397 fel) med en prick:

I försök 2 är testmaterialet däremot så omfattande att specialanpassning inte alls är en lika avgörande faktor för testresultatet. Vi har därför anledning att tro att de bästa knappsatserna från försöksserie 2 verkligen har egenskaper som gör dem lämpliga för inmatning av svensk text. Detta gäller dock bara det principiella utseendet och vi kan inte utesluta att små justeringar skulle få knappsatserna att presterå ännu bättre. Detta skall vi därför undersöka närmare i försöksserie 3.

### 6.6.3 Försöksserie 3

För att undersöka hur mindre justeringar av symbolplaceringen påverkar testresultatet väljer vi först ut de 1000 bästa knappsatserna från försöksserie 2:

![Figur 6.12: De 1000 bästa knappsatserna från försöksserie 2.](image)

Dessa knappsatser kommer att utgöra startdata för en iterationsalgoritm. I denna algoritm skapar vi i varje iteration nya knappsatser genom att byta plats på två symboler hos de 1000 ursprungliga knappsatserna. Detta görs tre gånger för varje knappsats vilket alltså genererar 3000 nya knappsatser. Testresultatet kontrolleras återigen med hjälp av testtext G och jämförs med
de tidigare resultaten. De nya och gamla knappsatserna slås därefter ihop och från den gruppen väljs återigen de 1000 bästa knappsatserna ut. Sedan upprepar vi hela förloppet det antal gånger vi önskar, i det här fallet 50.

Vi använder som vanligt matlab för att utföra våra försök och får följande resultat:

![Diagram](image)

Figur 6.13: Fördelningen av de 1000 knappsatser som använder i försöksserie 3, med avseende på antalet felaktiga symboler.


Vid en närmare undersökning visar det sig att våra små förändringar av en knappsats oftast ger sämre testresultat. En knappsats som gav kring 250 fel kunde efter byte av två symboler ge långt fler än 500 fel. Eftersom vi i varje iterationssteg väljer ut de 1000 bästa knappsatserna syns detta inte i vårt slutresultat men man kan konstatera att symbolplaceringen är en väldigt känslig faktor. Eftersom det vanligtvis blir sämre resultat efter ett byte är många av de 1000 knappsatser vi slutligen valde ut, desamma som vi startade med.

6.6.4 Försöksserie 4

Av de tusen knappsatser vi fick fram i försöksserie 3 väljer vi ut de 100 som gav bäst testresultat:

![Diagram](image1)

Figur 6.14: De 100 bästa knappsatserna från försöksserie 3.

Det visar sig att 58 av de 100 knappsatserna har symbolplaceringar som skiljer sig åt och det är dessa 58 vi ska använda i vår avslutande försöksserie.

För att verkligen gå till botten med vilken av dessa som är att föredra matar vi nu in testtext *H* som består av hela 100.000 tecken. Vi bedömer att det materialet är tillräckligt omfattande för att verkligen ge oss en god uppfattning om vilken knappsats som är att föredra i ett allmänt fall.

![Diagram](image2)

Figur 6.15: Antalet felaktiga symboler hos de 58 knappsatser som undersöktes i försöksserie 4. För att göra det mer överskådligt har resultaten sorterats i stigande ordning.
Den bästa knappsatsen i vår sista försöksserie ger 2709 felaktiga symboler för testtext G. Det är just denna knappsats som blir vårt slutgiltiga val.

6.7 Slutligt val av knappsats

Den knappsats vi slutligen bestämde oss för har följande utseende:

![Figur 6.16: Vårt slutliga val av knappsats.]

I den fortsatta texten är det underförstått att beteckningen Qtap nu refererar till ett system som är utrustat med denna knappsats.

6.7.1 Utvärdering

Vad kan vi då säga om den knappsatsen vi har valt? Ja, först och främst skall vi se hur den presterar jämfört med våra tidigare förslag. Vi matar därför in de fem testtexterna och får följande resultat:

<table>
<thead>
<tr>
<th>Testtext</th>
<th>Fel</th>
<th>Rätt</th>
<th>Felprocent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testtext A</td>
<td>2</td>
<td>124</td>
<td>1.6</td>
</tr>
<tr>
<td>Testtext B</td>
<td>12</td>
<td>184</td>
<td>6.6</td>
</tr>
<tr>
<td>Testtext C</td>
<td>45</td>
<td>1558</td>
<td>2.9</td>
</tr>
<tr>
<td>Testtext D</td>
<td>33</td>
<td>496</td>
<td>6.1</td>
</tr>
<tr>
<td>Testtext E</td>
<td>2</td>
<td>98</td>
<td>4.0</td>
</tr>
<tr>
<td>Totalt</td>
<td>94</td>
<td>2509</td>
<td>3.7</td>
</tr>
</tbody>
</table>

![Figur 6.17: Testresultat för vårt slutliga val av knappsats.]

47
En jämförelse med motsvarande testresultat för den traditionella och intuitiva knappsatsen ser vi i figur 6.18:

**Figur 6.18: En jämförelse mellan resultatet för våra tre knappsatser då testexterna A-E matas in.**

Speciellt skall vi återigen studera testtext A som med vår nya knappsats endast innehåller två felaktiga symboler:

**JAG VET EJ MED VILKA VAPEN DET BREDDDE VÅRLDSKRIGET KOMMER ATT UTKÄMPAS MEN DET FJÄRDE KOMMER ATT UTKÄMPAS MED PÅKAR OCH STENAR**

Så när som på ordet ”tredje” lyckas vårt system nu återge meddelandet helt korrekt. Resultatet är visserligen uppmuntrande men vi kan inte dra för långtgående slutsatser ifrån detta eftersom vi bara tittar på ett enskilt fall. Konsekvensen av det relativt begränsade material som de fem första testexterna utgör blir också synlig när vi jämför med den intuitiva knappsatsen. Vi kan notera att vi totalt sett inte får särskilt markant bättre testresultat. Till exempel ger den nya knappsatsen sämre testresultat för både testtext B och testtext C. Vi ska dock komma ihåg att vi valt ut knappsatsen utifrån ett väldigt omfattande testmaterial. Att andra knappsatser kan vara bättre i enskilda fall är inte oväntat. Till exempel är fullt möjligt att hitta knappsatser som klarar av att återge testtext A helt utan fel. För att få en mer rättvis jämförelse mellan hur de olika knappsatserna presterar matar vi in testtext G som ju består av 100.000 tecken.
Förenklad textinmatning på mobila enheter med hjälp av kontextbaserad språktolkning

Knappsats | Fel | Rätt | Felprocent
---|---|---|---
Traditionell | 5293 | 94707 | 5.3
Intuitiv | 4212 | 95788 | 4.2
Slutlig | 2709 | 97291 | 3.0

Figur 6.19: Resultatet för samtliga tre knappsatser då testtext $G$ matas in.

Här ser vi tydligare fördelarna med vårt val av knappsats. Utförligare information om hur vår slutliga knappsats presterar i olika fall återkommer vi till i nästa kapitel.

Tidigare hade vi några förslag på egenskaper som vi antog att en bra knappsats bör ha. Eftersom vi med datorns hjälp nu har en knappsats som verkliga ger bra testresultat är det intressant att se hur dessa egenskaper stämmer överens med denna. Först tittar vi på de sammanlagda sannolikheterna för varje knapp:

![Sannolikhetsfördelning mellan knappar](image)

Figur 6.20: Sannolikhetsfördelning mellan knapparna på vår slutliga knappsats.

Resultatet är något överraskande. Tydligen är sannolikhetsfördelningen mellan knapparna inte en så avgörande faktor som vi först antog. Fördelningen hos vår nya knappsats är ju minst lika avvikande som hos den traditionella. En sak som vi dock kan stå fast vid är att en bra knappsats inte
har råd att ”slösa” knappar som den traditionella knappsatsens gör genom att bara placera väldigt ovanliga bokstäver på knapp 9.


Som en avslutande jämförelse tittar vi på vinkelavståndet för den traditionella respektive den intuitiva knappsatsen:

<table>
<thead>
<tr>
<th>Knappsats</th>
<th>Vinkelavstånd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditionell</td>
<td>1.4647</td>
</tr>
<tr>
<td>Intuitiv</td>
<td>1.4694</td>
</tr>
<tr>
<td>Slutlig</td>
<td>1.4824</td>
</tr>
</tbody>
</table>

Figur 6.21: Vinkelavståndet för de olika knappsatserna.

Tittar vi tillbaka och jämför med de framslumpade knappsatserna i figur 6.8 ser vi att även den traditionella och den intuitiva knappsatsen har ett relativt högt värde på vinkelavståndet. Ingen kommer dock in närheten av vårt slutliga val knappsats.

6.8 Avslutning

Genom valet av knappsats har vi uppfyllt en stor del av målet för detta examensarbete. Ett textinmatningssystem som bygger på språkliga sannolikheter har utarbetats och en förändring av knappsatsens utseende har utretts. Vad som nu återstår är att se hur vårt system presterar jämfört med t9.
7 Utvärdering


7.1 Inledning

När vi nu har bestämt oss för vilken strategi och knappsats vi vill använda är det dags att jämföra vårt system med ett kommersiellt alternativ. Vi påminner om att vi valt att kalla vårt system Qtap. Som vi poängterat i avsnitt 5.1.5 kommer vi bara att utföra jämförelsen på en nivå som inte tar hänsyn till användarinteraktion. Av detta skäl kommer vi inte att jämföra vårt system med multitap eller Letterwise. Visserligen har vi i avsnitt 5.3.3 tagit fram grundprincipen för den strategi Letterwise utnyttjar men då systemet i så hög grad är beroende av att användaren korrigera eventuella fel efter hand, skulle en jämförelse bli väldigt orättvis. Därför har valet fallit på t9, som också är det överlägset mest spridda alternativet på dagens mobiltelefoner.

7.2 Förutsättningar

Innan vi påbörjar vår jämförelse skall vi gå igenom de förutsättningar som gäller för våra tester.

7.2.1 Utförande


7.2.2 Felmått


51
kommer ordfel emellanåt ge en rättvisare jämförelse mellan t9 och vårt system. Manuell inmatning innebär emellertid en del begränsningar.

### 7.2.3 Begränsningar


- När vi får en indikation på att ett ord saknas i ordlistan väljs det som systemet skrivit så långt som utdata. Därefter matas återstoden av ordet in som ett ”nytt ord”. Låt oss titta på ett kort exempel:

  Antag att vi vill mata in ordet **FERIEHUS**. På en traditionell knappssats innebär det att vi skall trycka in knappsekvensen **pptqpqut**. När vi försöker mata in den femte bokstaven kommer systemet emellertid berätta för oss att ordlistan inte innehåller något ord som motsvarar knappsekvensen. Utifrån de fyra första knapptryckningarna, **pptqa**, har systemet så långt gissat att vi vill skriva **DERI** och detta väljer vi därför som utdata. Därefter matar vi in återstoden av ordet, d.v.s. **pqut**. Enligt systemet är den mest troliga motsvarigheten till denna sekvens **EHUR**. Vårt försök att mata in **FERIEHUS** kommer alltså att ge **DERIEHUR**. På en symbolnivå kan detta emellanåt ge orättvist många fel för t9 eftersom systemet kanske hade lyckats bättre om vi till exempel hade delat upp det sammansatta ordet. Detta är ytterligare ett skäl till att vi har utvidgat vårt felmått. Då vi betraktar ordfel kommer sådana hårfall inte bli något problem. Ett saknat ord räknas helt enkelt som ett felaktigt ord och det blir oviktigt hur det såg ut i detalj.

- När en kollision inträffar och vi tvingas välja mellan olika möjliga utdata kommer vi genomgående välja det först föreslagna ordet. Vill vi skriva **HUS** trycker vi in knappsekvensen **qut**. Det mest sannolika ordet med denna knappsekvens är emellertid **HUR** som alltså blir vårt val av utdata.
7.3 Tester

Med dessa förutsättningar är vi nu redo att inleda vår jämförelse. Vi kommer att genomföra flera olika typer av tester som beskrivs i detalj nedan.

7.3.1 Texttext A-E

För att följa upp det vi gjort tidigare tittar vi på de fem testtexter som vi studerade i föregående kapitel. Vi mata in dessa texter med t9 och jämför resultatet med det vi erhöll för det egna systemet. Vi genomför också kompletterande inmatningar för att jämföra på fel på ordnivå.

![Figur 7.1: Antalet felaktiga symboler då testtext A-E matas in med t9 respektive Qtap.](image1)

![Figur 7.2: Antalet felaktiga ord då testtext A-E matas in med t9 respektive Qtap.](image2)
Vi kan konstatera att de totala testresultaten är relativt likartade. Qtap har ett litet försprång då vi betraktar antalet felaktiga symboler medan t9 är något bättre när det gäller ord. Att t9 lyckas bättre på ordnivå än på symbolnivå är inte förvånande. Då ett ord saknas i ordlistan löper t9 en stor risk att ge många felaktiga symboler med våra förutsättningar i avsnitt 7.2.3. På grund av detta kommer de felaktiga symbolerna ofta att hopa sig och förekomma i flertal i ett och samma ord. Hos Qtap får vi däremot en mer jämn utspridning av felaktiga symboler.

Vi har redan tidigare konstaterat att de fem testtexterna utgör ett ganska begränsat testmaterial. Därför skall vi i vårt nästa försök se hur de två systemen klarar sig när vi matar in en längre text.

### 7.3.2 Lång text


### 7.3.3 Typiska textmeddelanden

Vår nästa undersökning kommer att handla om det som idag är det främsta användningsområdet för mobiltelefonens textinmatningssystem, nämligen sms. Vårt testunderlag består av 100 insamlade sms av varierande längd. Insamlingen har skett genom att tio olika personer i åldrarna 16-58 år har fått skriva ned tio sms som fanns sparade på deras mobiltelefon. Dessa har sedan matas med både t9 och vårt nya system och antalet felaktiga symboler respektive ord har noterats.
Resultaten blir som synes väldigt snarlika. Totalt bestod de 100 textmeddelandena av 10240 tecken fördelade på 2052 ord. Omfattningen är alltså likvärdig med den långa text vi använde i föregående försök och vi måste fråga oss varför inte testresultaten också är liknande. Ja, att t9 presterar något sämre är inte förvånande. Typisk sms-text avviker i allmänhet mer från ”korrekt svenska” än romanprosa, vilket gör att fler ord kommer att saknas i ordlistan. Det kraftigt försämrade resultatet för Qtap har en liknande förklaring. Textmeddelandena innehåller till exempel flera akronymer och slanguttryck som inte finns representerade i vårt statistikmaterial. Till exempel visar det sig mkt och oxå, som inte sällan förekommer i sms-text, inte finns representerade någon enda gång i PAROLE. Andra exempel på problematiska förkortningar och slanguttryck som förekom i testmaterialet är iaf, mezza, ql(kul), sms, lixom och kramiz.

### 7.3.4 Långa ord

I detta försök skall vi studera hur de båda systemen klarar av inmatning av långa ord. På förhand misstänker vi att t9 kan få problem med dessa. Många längre ord är sammansatta och eftersom det finns så många varianter av ordbildningar kan alla dessa inte finnas lagrade i ordlistan. Vi bör alltså få ett ganska stort antal ord som t9 inte lyckas återge korrekt. Försöksunderlaget hämtar vi från ett långt textavsnitt i PAROLE. Vi kommer att välja ut ord som består av minst tio bokstäver och se till att inga upprepningar förekommer. De 100 första orden, bestående av totalt 1169 tecken, matar vi sedan in med både t9 och Qtap. Resultatet blir följande:
Förenklad textinmatning på mobila enheter med hjälp av kontextbaserad språktolkning

Figur 7.6: Sammanlagt testresultat för t9 och Qtap efter inmatning av 100 ord bestående av minst tio tecken.

Precis som vi misstänkte har Qtap en fördel över t9. Övertaget vad gäller symbolfel är väldigt markant men det är framförallt jämförelsen av antalet felaktiga ord som är intressant i detta försök. Eftersom vi inte tillåter någon interaktion mellan system och användare kommer t9 nämligen få stora problem när inmatningen av ett ord misslyckas. Om detta sker tidigt under inmatningen riskerar återstoden av ordet att innehålla många fel. Även Qtap har vissa problem med långa ord. Detta understryks av att drygt 9 % av symbolerna blir fel i vårt försök med långa ord. Det kan vi till exempel jämföra med vårt tidigare försök med lång text där motsvarande siffra var ca 2.6%. Vi ser också att inmatningen av en ganska stor andel av de 100 orden misslyckas för båda systemen.

7.3.5 Egennamn

Egennamn av olika slag är vanligt förekommande i typiska textmeddelanden. Bland annat av den anledningen har t9 inkluderat väldigt många för- och efternamn i sin ordlista. Antalet namn är ju emellertid enormt så problem med namninmatning kommer ändå att uppstå. Vi ska genomföra ett försök där vi mäter in namn på personer, städer och länder med t9 och Qtap. Testmaterialet består av 50 för- och efternamn som är hämtade från en slumpvis utvald lista över studenter vid Linköpings Universitet. Sammanlagt handlar det alltså om 100 inmatningar som totalt består av 714 tecken. Resultatet visas i nedanstående figur:
Vi ser att Qtap får stora problem med namnen. Anledningen till detta står att finna i namnens struktur. Ofta har namn en språklig struktur som i vissa avseenden avviker från normalt svenskt skriftspråk. Vi har till exempel många namn av utländsk härkomst, som innehåller bokstavsmönster som inte är särskilt frekvent i svenskans. T9 som har många av de vanligare namnen i sin ordlista klarar sig däremot mycket bättre.
8 Resultat

I det här kapitlet sammanfattas alla resultat och vilka slutsatser som kan dras utifrån detta redovisas. Därefter beskrivs och kommenteras tänkbara utvidgningar för arbetet. Till sist följer en kort utblick om framtidens textinmatningssystem.

8.1 Slutsatser

Vi har tagit fram principerna för ett textinmatningssystem till mobiltelefoner som utnyttjar sannolikheter för mönster i språket. Vi har också funnit ett nytt utseende på knappsatsen som ger systemet bättre möjlighet att avgöra önska inmatning från användaren. Därefter har vi genomfört ett antal olika tester för att jämföra vårt framtagna system, Qtap, med t9.

Sammanfattningsvis kan vi konstatera att testresultaten är relativt likvärdiga totalt sett, även om det ena systemet kan ha fördelar jämfört med det andra vid vissa typer av inmatningar. Inmatning av sms är den överlägsen vanligaste tillämpningen för mobila textinmatningssystem idag och där uppvisar de båda systemen väldigt lika testresultat. Frågan vilket system som är "bäst" får nog därmed anses obesvarad. Dessutom skall vi komma ihåg att vi jämfört systemen på en nivå som inte tillåter samspel med användaren. För att riktigt gå till botten med vilket system som är att föredra i allmänhet, skulle vi vara tvungen att utvidga Qtap till att även omfatta användarinteraktion. Först då kan jämförelsen bli riktigt rättvis. En sådan utvidgning ingår som vi konstaterat tidigare inte i arbetsuppgifterna för detta examensarbete.

Detta gör att vi endast betraktar testresultaten som vägledande. Slutsatsen vi drar är först och främst att resultaten är tillräckligt goda för att en vidare studie av de principer som Qtap bygger på, är motiverad. Ingenting som framkommit i detta arbete tyder på att viterbiavkodning bör förkastas som strategi vid textinmatning.

8.2 Fortsatt arbete

8.2.1 Användargränssnitt

Den mest uppenbara utvidgningen av vårt arbete är att utveckla ett användargränssnitt för Qtap. Först då kan en helt rättvis jämförelse med andra kommersiella system genomföras. Idag används ofta ett mått som kallas KSPC när man jämför olika textinmatningssystem. KSPC står för keystrokes per character och anger hur många knapptryckningar det i genomsnitt krävs för att mata in en symbol med en given textinmatningsteknik. För multitap och t9 är till exempel KSPC = 2.0342 respektive 1.0072 [10]. I detta mått är även de knapptryckningar användaren
gör för att korrigera eventuella fel medräknade. Att t9:s värde överstiger 1 beror på att vi emellanåt måste bläddra och välja mellan olika alternativ till en given knappsekvens. (Värdet är framtaget under den generösa förutsättningen att inga ord som matas in saknas i ordlistan). KSPC-måttet bygger alltså på att användaren på något sätt kan se till att utdata slutligen motsvarar den önskade inmatningen. Denna möjlighet har vi ännu inte hos Qtap och det vore förstås önskvärt att införa detta så vi kan göra en beräkning av systemets KSPC-värde.


Som synes är det en ganska grannlaga uppgift att utveckla ett användargränssnitt för Qtap. Bland annat måste en lämplig sekvenslängd och ett smidigt sätt att korrigera fel bestämmas. Om ett bra användargränssnitt kan hittas synes dock utsikterna för att systemet skall klara sig bra i en jämförelse med kommersiella alternativ, mycket goda.

8.2.2 Förbättring av modellen

8.2.3 Undersökning av hårdvarukrav

En annan faktor som sätter gränser för vilka modeller som kan användas och hur mycket statistik som kan lagras är minneskapacitet. I detta arbete har vi inte gjort någon undersökning av vilka hårdvarukrav Qtap ställer. Om vi i slutändan vill utvärdera Qtap i ett kommersiellt alternativ är denna faktor av yttersta vikt. Givet dagens snabba utveckling och det faktum att storleken på mobiltelefoner inte kan krympa särskilt mycket mer, är det dock rimligt att tro att de krav Qtap ställer kan tillgodoses.

8.2.4 Utveckling av knappsatsern


Vi kan tänka oss att beräkningen av det genomsnittliga vinkelavståndet kan vara något missvisande och möjligen går det hitta ett bättre sätt att ta hänsyn till vinkelavståndet. Det skulle till exempel kunna vara önskvärt att studera större omgivningar till varje bokstav.
8.3 Framtidens textinmatningssystem


komplett knappslats, nämligen felskrivningar och långsam inmatning. Främst beror detta på att trögheten i knapparna har valts så att feltryckningar kan undvikas i högst möjliga mån.

8.4 Avslutning

Referenser

Förenklat textinmatning på mobila enheter
med hjälp av kontextbaserad språktolkning

A Språkstatistiska tabeller

A.1 Bokstavsfrekvenser

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Antal belägg</th>
<th>Frekvens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2098225</td>
<td>0,081</td>
</tr>
<tr>
<td>A</td>
<td>936460</td>
<td>0,081</td>
</tr>
<tr>
<td>E</td>
<td>894010</td>
<td>0,077</td>
</tr>
<tr>
<td>N</td>
<td>836229</td>
<td>0,072</td>
</tr>
<tr>
<td>T</td>
<td>812742</td>
<td>0,070</td>
</tr>
<tr>
<td>R</td>
<td>708343</td>
<td>0,061</td>
</tr>
<tr>
<td>S</td>
<td>551273</td>
<td>0,047</td>
</tr>
<tr>
<td>D</td>
<td>479683</td>
<td>0,041</td>
</tr>
<tr>
<td>L</td>
<td>470688</td>
<td>0,040</td>
</tr>
<tr>
<td>I</td>
<td>460103</td>
<td>0,040</td>
</tr>
<tr>
<td>O</td>
<td>398952</td>
<td>0,034</td>
</tr>
<tr>
<td>G</td>
<td>343264</td>
<td>0,030</td>
</tr>
<tr>
<td>M</td>
<td>329858</td>
<td>0,028</td>
</tr>
<tr>
<td>H</td>
<td>297777</td>
<td>0,026</td>
</tr>
<tr>
<td>K</td>
<td>293750</td>
<td>0,025</td>
</tr>
<tr>
<td>V</td>
<td>235666</td>
<td>0,020</td>
</tr>
<tr>
<td>A</td>
<td>209418</td>
<td>0,018</td>
</tr>
<tr>
<td>Å</td>
<td>187795</td>
<td>0,016</td>
</tr>
<tr>
<td>U</td>
<td>182532</td>
<td>0,016</td>
</tr>
<tr>
<td>F</td>
<td>168740</td>
<td>0,015</td>
</tr>
<tr>
<td>P</td>
<td>153415</td>
<td>0,013</td>
</tr>
<tr>
<td>Ö</td>
<td>143369</td>
<td>0,012</td>
</tr>
<tr>
<td>C</td>
<td>140390</td>
<td>0,012</td>
</tr>
<tr>
<td>B</td>
<td>119392</td>
<td>0,010</td>
</tr>
<tr>
<td>J</td>
<td>117540</td>
<td>0,010</td>
</tr>
<tr>
<td>Y</td>
<td>49150</td>
<td>0,004</td>
</tr>
<tr>
<td>X</td>
<td>7419</td>
<td>0,001</td>
</tr>
<tr>
<td>W</td>
<td>1710</td>
<td>0,0001</td>
</tr>
<tr>
<td>Z</td>
<td>1346</td>
<td>0,0001</td>
</tr>
<tr>
<td>Q</td>
<td>184</td>
<td>0,0002</td>
</tr>
</tbody>
</table>
### A.2 Bigramfrekvenser

Tabellen visar de 100 vanligaste bigrammen i PAROLE med tillhörande frekvenser:

<table>
<thead>
<tr>
<th>Nr</th>
<th>Bigram</th>
<th>Frekvens</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N</td>
<td>0.1112</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>0.1050</td>
</tr>
<tr>
<td>3</td>
<td>S</td>
<td>0.0889</td>
</tr>
<tr>
<td>4</td>
<td>R</td>
<td>0.0879</td>
</tr>
<tr>
<td>5</td>
<td>DE</td>
<td>0.0838</td>
</tr>
<tr>
<td>6</td>
<td>E</td>
<td>0.0791</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>0.0744</td>
</tr>
<tr>
<td>8</td>
<td>EN</td>
<td>0.0698</td>
</tr>
<tr>
<td>9</td>
<td>H</td>
<td>0.0631</td>
</tr>
<tr>
<td>10</td>
<td>D</td>
<td>0.0578</td>
</tr>
<tr>
<td>11</td>
<td>AN</td>
<td>0.0567</td>
</tr>
<tr>
<td>12</td>
<td>AR</td>
<td>0.0524</td>
</tr>
<tr>
<td>13</td>
<td>ET</td>
<td>0.0463</td>
</tr>
<tr>
<td>14</td>
<td>M</td>
<td>0.0432</td>
</tr>
<tr>
<td>15</td>
<td>ER</td>
<td>0.0423</td>
</tr>
<tr>
<td>16</td>
<td>O</td>
<td>0.0422</td>
</tr>
<tr>
<td>17</td>
<td>V</td>
<td>0.0421</td>
</tr>
<tr>
<td>18</td>
<td>F</td>
<td>0.0406</td>
</tr>
<tr>
<td>19</td>
<td>G</td>
<td>0.0398</td>
</tr>
<tr>
<td>20</td>
<td>IN</td>
<td>0.0395</td>
</tr>
<tr>
<td>21</td>
<td>TT</td>
<td>0.0381</td>
</tr>
<tr>
<td>22</td>
<td>TE</td>
<td>0.0376</td>
</tr>
<tr>
<td>23</td>
<td>A</td>
<td>0.0374</td>
</tr>
<tr>
<td>24</td>
<td>I</td>
<td>0.0370</td>
</tr>
<tr>
<td>25</td>
<td>LL</td>
<td>0.0347</td>
</tr>
<tr>
<td>26</td>
<td>ST</td>
<td>0.0335</td>
</tr>
<tr>
<td>27</td>
<td>OC</td>
<td>0.0307</td>
</tr>
<tr>
<td>28</td>
<td>M</td>
<td>0.0305</td>
</tr>
<tr>
<td>29</td>
<td>T</td>
<td>0.0305</td>
</tr>
<tr>
<td>30</td>
<td>HA</td>
<td>0.0303</td>
</tr>
<tr>
<td>31</td>
<td>OM</td>
<td>0.0303</td>
</tr>
<tr>
<td>32</td>
<td>RA</td>
<td>0.0297</td>
</tr>
<tr>
<td>33</td>
<td>CH</td>
<td>0.0292</td>
</tr>
<tr>
<td>34</td>
<td>AD</td>
<td>0.0283</td>
</tr>
<tr>
<td>35</td>
<td>H</td>
<td>0.0282</td>
</tr>
<tr>
<td>36</td>
<td>B</td>
<td>0.0281</td>
</tr>
<tr>
<td>37</td>
<td>AT</td>
<td>0.0280</td>
</tr>
<tr>
<td>38</td>
<td>Ä</td>
<td>0.0273</td>
</tr>
<tr>
<td>39</td>
<td>K</td>
<td>0.0268</td>
</tr>
<tr>
<td>40</td>
<td>TA</td>
<td>0.0261</td>
</tr>
<tr>
<td>41</td>
<td>VA</td>
<td>0.0258</td>
</tr>
<tr>
<td>42</td>
<td>OR</td>
<td>0.0257</td>
</tr>
<tr>
<td>43</td>
<td>D</td>
<td>0.0256</td>
</tr>
<tr>
<td>44</td>
<td>E</td>
<td>0.0248</td>
</tr>
<tr>
<td>45</td>
<td>IG</td>
<td>0.0246</td>
</tr>
<tr>
<td>46</td>
<td>ND</td>
<td>0.0244</td>
</tr>
<tr>
<td>47</td>
<td>S</td>
<td>0.0243</td>
</tr>
<tr>
<td>48</td>
<td>NA</td>
<td>0.0237</td>
</tr>
<tr>
<td>49</td>
<td>ME</td>
<td>0.0235</td>
</tr>
<tr>
<td>50</td>
<td>KA</td>
<td>0.0234</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nr</th>
<th>Bigram</th>
<th>Frekvens</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>AR</td>
<td>0.0231</td>
</tr>
<tr>
<td>52</td>
<td>SK</td>
<td>0.0219</td>
</tr>
<tr>
<td>53</td>
<td>NG</td>
<td>0.0214</td>
</tr>
<tr>
<td>54</td>
<td>P</td>
<td>0.0211</td>
</tr>
<tr>
<td>55</td>
<td>LA</td>
<td>0.0210</td>
</tr>
<tr>
<td>56</td>
<td>I</td>
<td>0.0210</td>
</tr>
<tr>
<td>57</td>
<td>ON</td>
<td>0.0206</td>
</tr>
<tr>
<td>58</td>
<td>TI</td>
<td>0.0205</td>
</tr>
<tr>
<td>59</td>
<td>J</td>
<td>0.0203</td>
</tr>
<tr>
<td>60</td>
<td>AG</td>
<td>0.0202</td>
</tr>
<tr>
<td>61</td>
<td>N</td>
<td>0.0201</td>
</tr>
<tr>
<td>62</td>
<td>LI</td>
<td>0.0197</td>
</tr>
<tr>
<td>63</td>
<td>L</td>
<td>0.0195</td>
</tr>
<tr>
<td>64</td>
<td>G</td>
<td>0.0194</td>
</tr>
<tr>
<td>65</td>
<td>FO</td>
<td>0.0193</td>
</tr>
<tr>
<td>66</td>
<td>NT</td>
<td>0.0190</td>
</tr>
<tr>
<td>67</td>
<td>OR</td>
<td>0.0185</td>
</tr>
<tr>
<td>68</td>
<td>RE</td>
<td>0.0184</td>
</tr>
<tr>
<td>69</td>
<td>LE</td>
<td>0.0180</td>
</tr>
<tr>
<td>70</td>
<td>JA</td>
<td>0.0179</td>
</tr>
<tr>
<td>71</td>
<td>GE</td>
<td>0.0177</td>
</tr>
<tr>
<td>72</td>
<td>IL</td>
<td>0.0162</td>
</tr>
<tr>
<td>73</td>
<td>AN</td>
<td>0.0162</td>
</tr>
<tr>
<td>74</td>
<td>SA</td>
<td>0.0160</td>
</tr>
<tr>
<td>75</td>
<td>L</td>
<td>0.0159</td>
</tr>
<tr>
<td>76</td>
<td>CK</td>
<td>0.0156</td>
</tr>
<tr>
<td>77</td>
<td>AL</td>
<td>0.0153</td>
</tr>
<tr>
<td>78</td>
<td>NN</td>
<td>0.0153</td>
</tr>
<tr>
<td>79</td>
<td>VI</td>
<td>0.0152</td>
</tr>
<tr>
<td>80</td>
<td>SO</td>
<td>0.0152</td>
</tr>
<tr>
<td>81</td>
<td>NS</td>
<td>0.0148</td>
</tr>
<tr>
<td>82</td>
<td>MA</td>
<td>0.0148</td>
</tr>
<tr>
<td>83</td>
<td>GA</td>
<td>0.0143</td>
</tr>
<tr>
<td>84</td>
<td>ED</td>
<td>0.0141</td>
</tr>
<tr>
<td>85</td>
<td>NE</td>
<td>0.0136</td>
</tr>
<tr>
<td>86</td>
<td>HO</td>
<td>0.0136</td>
</tr>
<tr>
<td>87</td>
<td>U</td>
<td>0.0135</td>
</tr>
<tr>
<td>88</td>
<td>RI</td>
<td>0.0135</td>
</tr>
<tr>
<td>89</td>
<td>SI</td>
<td>0.0132</td>
</tr>
<tr>
<td>90</td>
<td>EL</td>
<td>0.0130</td>
</tr>
<tr>
<td>91</td>
<td>PÅ</td>
<td>0.0125</td>
</tr>
<tr>
<td>92</td>
<td>DA</td>
<td>0.0121</td>
</tr>
<tr>
<td>93</td>
<td>R</td>
<td>0.0121</td>
</tr>
<tr>
<td>94</td>
<td>Å</td>
<td>0.0120</td>
</tr>
<tr>
<td>95</td>
<td>SE</td>
<td>0.0118</td>
</tr>
<tr>
<td>96</td>
<td>KE</td>
<td>0.0116</td>
</tr>
<tr>
<td>97</td>
<td>IT</td>
<td>0.0112</td>
</tr>
<tr>
<td>98</td>
<td>HE</td>
<td>0.0109</td>
</tr>
<tr>
<td>99</td>
<td>VE</td>
<td>0.0108</td>
</tr>
<tr>
<td>100</td>
<td>UN</td>
<td>0.0106</td>
</tr>
</tbody>
</table>
### A.3 Trigramfrekvenser

Tabellen visar de 100 vanligaste trigrammen i PAROLE med tillhörande frekvenser:

<table>
<thead>
<tr>
<th>Nr</th>
<th>Trigram</th>
<th>Frekvens</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EN</td>
<td>0.0208</td>
</tr>
<tr>
<td>2</td>
<td>DE</td>
<td>0.0150</td>
</tr>
<tr>
<td>3</td>
<td>DE</td>
<td>0.0139</td>
</tr>
<tr>
<td>4</td>
<td>ET</td>
<td>0.0133</td>
</tr>
<tr>
<td>5</td>
<td>AN</td>
<td>0.0117</td>
</tr>
<tr>
<td>6</td>
<td>OC</td>
<td>0.0113</td>
</tr>
<tr>
<td>7</td>
<td>HA</td>
<td>0.0111</td>
</tr>
<tr>
<td>8</td>
<td>CH</td>
<td>0.0108</td>
</tr>
<tr>
<td>9</td>
<td>OCH</td>
<td>0.0107</td>
</tr>
<tr>
<td>10</td>
<td>TT</td>
<td>0.0101</td>
</tr>
<tr>
<td>11</td>
<td>AR</td>
<td>0.0100</td>
</tr>
<tr>
<td>12</td>
<td>OM</td>
<td>0.0096</td>
</tr>
<tr>
<td>13</td>
<td>ER</td>
<td>0.0095</td>
</tr>
<tr>
<td>14</td>
<td>ADE</td>
<td>0.0084</td>
</tr>
<tr>
<td>15</td>
<td>DET</td>
<td>0.0083</td>
</tr>
<tr>
<td>16</td>
<td>VA</td>
<td>0.0076</td>
</tr>
<tr>
<td>17</td>
<td>TE</td>
<td>0.0075</td>
</tr>
<tr>
<td>18</td>
<td>ATT</td>
<td>0.0073</td>
</tr>
<tr>
<td>19</td>
<td>ME</td>
<td>0.0073</td>
</tr>
<tr>
<td>20</td>
<td>IN</td>
<td>0.0069</td>
</tr>
<tr>
<td>21</td>
<td>FÖR</td>
<td>0.0069</td>
</tr>
<tr>
<td>22</td>
<td>VAR</td>
<td>0.0068</td>
</tr>
<tr>
<td>23</td>
<td>HAN</td>
<td>0.0065</td>
</tr>
<tr>
<td>24</td>
<td>FÖ</td>
<td>0.0063</td>
</tr>
<tr>
<td>25</td>
<td>AR</td>
<td>0.0059</td>
</tr>
<tr>
<td>26</td>
<td>N S</td>
<td>0.0059</td>
</tr>
<tr>
<td>27</td>
<td>AT</td>
<td>0.0058</td>
</tr>
<tr>
<td>28</td>
<td>JA</td>
<td>0.0058</td>
</tr>
<tr>
<td>29</td>
<td>AG</td>
<td>0.0058</td>
</tr>
<tr>
<td>30</td>
<td>J</td>
<td>0.0057</td>
</tr>
<tr>
<td>31</td>
<td>JAG</td>
<td>0.0052</td>
</tr>
<tr>
<td>32</td>
<td>EN</td>
<td>0.0051</td>
</tr>
<tr>
<td>33</td>
<td>SO</td>
<td>0.0051</td>
</tr>
<tr>
<td>34</td>
<td>NDE</td>
<td>0.0051</td>
</tr>
<tr>
<td>35</td>
<td>HO</td>
<td>0.0049</td>
</tr>
<tr>
<td>36</td>
<td>NA</td>
<td>0.0048</td>
</tr>
<tr>
<td>37</td>
<td>T S</td>
<td>0.0048</td>
</tr>
<tr>
<td>38</td>
<td>RA</td>
<td>0.0048</td>
</tr>
<tr>
<td>39</td>
<td>ILL</td>
<td>0.0048</td>
</tr>
<tr>
<td>40</td>
<td>ON</td>
<td>0.0047</td>
</tr>
<tr>
<td>41</td>
<td>PÅ</td>
<td>0.0047</td>
</tr>
<tr>
<td>42</td>
<td>SOM</td>
<td>0.0047</td>
</tr>
<tr>
<td>43</td>
<td>PA</td>
<td>0.0046</td>
</tr>
<tr>
<td>44</td>
<td>SK</td>
<td>0.0046</td>
</tr>
<tr>
<td>45</td>
<td>NTE</td>
<td>0.0045</td>
</tr>
<tr>
<td>46</td>
<td>INT</td>
<td>0.0045</td>
</tr>
<tr>
<td>47</td>
<td>IG</td>
<td>0.0044</td>
</tr>
<tr>
<td>48</td>
<td>VI</td>
<td>0.0044</td>
</tr>
<tr>
<td>49</td>
<td>HON</td>
<td>0.0043</td>
</tr>
<tr>
<td>50</td>
<td>ING</td>
<td>0.0043</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nr</th>
<th>Trigram</th>
<th>Frekvens</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>DEN</td>
<td>0.0042</td>
</tr>
<tr>
<td>52</td>
<td>_ST</td>
<td>0.0042</td>
</tr>
<tr>
<td>53</td>
<td>_TI</td>
<td>0.0041</td>
</tr>
<tr>
<td>54</td>
<td>SI</td>
<td>0.0041</td>
</tr>
<tr>
<td>55</td>
<td>LL</td>
<td>0.0041</td>
</tr>
<tr>
<td>56</td>
<td>R D</td>
<td>0.0040</td>
</tr>
<tr>
<td>57</td>
<td>AND</td>
<td>0.0040</td>
</tr>
<tr>
<td>58</td>
<td>GEN</td>
<td>0.0040</td>
</tr>
<tr>
<td>59</td>
<td>OR</td>
<td>0.0038</td>
</tr>
<tr>
<td>60</td>
<td>R S</td>
<td>0.0037</td>
</tr>
<tr>
<td>61</td>
<td>E S</td>
<td>0.0037</td>
</tr>
<tr>
<td>62</td>
<td>A S</td>
<td>0.0037</td>
</tr>
<tr>
<td>63</td>
<td>LLE</td>
<td>0.0036</td>
</tr>
<tr>
<td>64</td>
<td>N H</td>
<td>0.0036</td>
</tr>
<tr>
<td>65</td>
<td>TA</td>
<td>0.0036</td>
</tr>
<tr>
<td>66</td>
<td>MED</td>
<td>0.0036</td>
</tr>
<tr>
<td>67</td>
<td>T H</td>
<td>0.0035</td>
</tr>
<tr>
<td>68</td>
<td>ED</td>
<td>0.0035</td>
</tr>
<tr>
<td>69</td>
<td>TIL</td>
<td>0.0034</td>
</tr>
<tr>
<td>70</td>
<td>E H</td>
<td>0.0034</td>
</tr>
<tr>
<td>71</td>
<td>SÅ</td>
<td>0.0033</td>
</tr>
<tr>
<td>72</td>
<td>SA</td>
<td>0.0032</td>
</tr>
<tr>
<td>73</td>
<td>T D</td>
<td>0.0032</td>
</tr>
<tr>
<td>74</td>
<td>TER</td>
<td>0.0031</td>
</tr>
<tr>
<td>75</td>
<td>ALL</td>
<td>0.0031</td>
</tr>
<tr>
<td>76</td>
<td>LIG</td>
<td>0.0031</td>
</tr>
<tr>
<td>77</td>
<td>HE</td>
<td>0.0031</td>
</tr>
<tr>
<td>78</td>
<td>SÅ</td>
<td>0.0031</td>
</tr>
<tr>
<td>79</td>
<td>MEN</td>
<td>0.0030</td>
</tr>
<tr>
<td>80</td>
<td>R H</td>
<td>0.0030</td>
</tr>
<tr>
<td>81</td>
<td>T V</td>
<td>0.0030</td>
</tr>
<tr>
<td>82</td>
<td>AR</td>
<td>0.0030</td>
</tr>
<tr>
<td>83</td>
<td>N V</td>
<td>0.0030</td>
</tr>
<tr>
<td>84</td>
<td>N D</td>
<td>0.0030</td>
</tr>
<tr>
<td>85</td>
<td>KÅ</td>
<td>0.0029</td>
</tr>
<tr>
<td>86</td>
<td>N O</td>
<td>0.0028</td>
</tr>
<tr>
<td>87</td>
<td>OM</td>
<td>0.0027</td>
</tr>
<tr>
<td>88</td>
<td>MA</td>
<td>0.0027</td>
</tr>
<tr>
<td>89</td>
<td>AL</td>
<td>0.0027</td>
</tr>
<tr>
<td>90</td>
<td>FR</td>
<td>0.0026</td>
</tr>
<tr>
<td>91</td>
<td>RNA</td>
<td>0.0026</td>
</tr>
<tr>
<td>92</td>
<td>SKA</td>
<td>0.0026</td>
</tr>
<tr>
<td>93</td>
<td>LA</td>
<td>0.0026</td>
</tr>
<tr>
<td>94</td>
<td>T O</td>
<td>0.0026</td>
</tr>
<tr>
<td>95</td>
<td>N F</td>
<td>0.0025</td>
</tr>
<tr>
<td>96</td>
<td>AV</td>
<td>0.0025</td>
</tr>
<tr>
<td>97</td>
<td>STE</td>
<td>0.0025</td>
</tr>
<tr>
<td>98</td>
<td>GT</td>
<td>0.0025</td>
</tr>
<tr>
<td>99</td>
<td>ETÅ</td>
<td>0.0024</td>
</tr>
<tr>
<td>100</td>
<td>T F</td>
<td>0.0024</td>
</tr>
</tbody>
</table>
### A.4 Tetragramfrekvenser

Tabellen visar de 100 vanligaste tetragrammen i PAROLE med tillhörande frekvenser:

<table>
<thead>
<tr>
<th>Nr</th>
<th>Tetragram</th>
<th>Frekvens</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>_OCH</td>
<td>0.0356</td>
</tr>
<tr>
<td>2</td>
<td>OCH_</td>
<td>0.0356</td>
</tr>
<tr>
<td>3</td>
<td>ADE_</td>
<td>0.0256</td>
</tr>
<tr>
<td>4</td>
<td>DET_</td>
<td>0.0248</td>
</tr>
<tr>
<td>5</td>
<td>ATT_</td>
<td>0.0244</td>
</tr>
<tr>
<td>6</td>
<td>HAN_</td>
<td>0.0243</td>
</tr>
<tr>
<td>7</td>
<td>DET_</td>
<td>0.0237</td>
</tr>
<tr>
<td>8</td>
<td>JAG_</td>
<td>0.0235</td>
</tr>
<tr>
<td>9</td>
<td>JAG_</td>
<td>0.0234</td>
</tr>
<tr>
<td>10</td>
<td>ATT_</td>
<td>0.0230</td>
</tr>
<tr>
<td>11</td>
<td>HAN_</td>
<td>0.0205</td>
</tr>
<tr>
<td>12</td>
<td>FÖR_</td>
<td>0.0205</td>
</tr>
<tr>
<td>13</td>
<td>VAR_</td>
<td>0.0200</td>
</tr>
<tr>
<td>14</td>
<td>HON_</td>
<td>0.0192</td>
</tr>
<tr>
<td>15</td>
<td>INT_</td>
<td>0.0162</td>
</tr>
<tr>
<td>16</td>
<td>SOM_</td>
<td>0.0161</td>
</tr>
<tr>
<td>17</td>
<td>SOM_</td>
<td>0.0160</td>
</tr>
<tr>
<td>18</td>
<td>INTE_</td>
<td>0.0158</td>
</tr>
<tr>
<td>19</td>
<td>PÅ_</td>
<td>0.0155</td>
</tr>
<tr>
<td>20</td>
<td>NTE_</td>
<td>0.0155</td>
</tr>
<tr>
<td>21</td>
<td>HON_</td>
<td>0.0149</td>
</tr>
<tr>
<td>22</td>
<td>VAR_</td>
<td>0.0141</td>
</tr>
<tr>
<td>23</td>
<td><em>EN</em></td>
<td>0.0139</td>
</tr>
<tr>
<td>24</td>
<td>MED_</td>
<td>0.0131</td>
</tr>
<tr>
<td>25</td>
<td>NDE_</td>
<td>0.0125</td>
</tr>
<tr>
<td>26</td>
<td>DEN_</td>
<td>0.0119</td>
</tr>
<tr>
<td>27</td>
<td>MED_</td>
<td>0.0119</td>
</tr>
<tr>
<td>28</td>
<td>FOR_</td>
<td>0.0117</td>
</tr>
<tr>
<td>29</td>
<td>TILL_</td>
<td>0.0108</td>
</tr>
<tr>
<td>30</td>
<td>TIL_</td>
<td>0.0103</td>
</tr>
<tr>
<td>31</td>
<td>AR_</td>
<td>0.0101</td>
</tr>
<tr>
<td>32</td>
<td>ILL_</td>
<td>0.0098</td>
</tr>
<tr>
<td>33</td>
<td>OM_</td>
<td>0.0094</td>
</tr>
<tr>
<td>34</td>
<td>GEN_</td>
<td>0.0092</td>
</tr>
<tr>
<td>35</td>
<td>MEN_</td>
<td>0.0090</td>
</tr>
<tr>
<td>36</td>
<td>EN S</td>
<td>0.0087</td>
</tr>
<tr>
<td>37</td>
<td>MEN_</td>
<td>0.0086</td>
</tr>
<tr>
<td>38</td>
<td>DEN_</td>
<td>0.0086</td>
</tr>
<tr>
<td>39</td>
<td>R DE_</td>
<td>0.0085</td>
</tr>
<tr>
<td>40</td>
<td>SÅ_</td>
<td>0.0081</td>
</tr>
<tr>
<td>41</td>
<td>ANDE_</td>
<td>0.0080</td>
</tr>
<tr>
<td>42</td>
<td>DU_</td>
<td>0.0075</td>
</tr>
<tr>
<td>43</td>
<td>RNA_</td>
<td>0.0073</td>
</tr>
<tr>
<td>44</td>
<td>HADE_</td>
<td>0.0072</td>
</tr>
<tr>
<td>45</td>
<td>HAD_</td>
<td>0.0072</td>
</tr>
<tr>
<td>46</td>
<td>DE_</td>
<td>0.0072</td>
</tr>
<tr>
<td>47</td>
<td>LLE_</td>
<td>0.0071</td>
</tr>
<tr>
<td>48</td>
<td>IGT_</td>
<td>0.0071</td>
</tr>
<tr>
<td>49</td>
<td>DE H_</td>
<td>0.0069</td>
</tr>
<tr>
<td>50</td>
<td>T DE_</td>
<td>0.0068</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nr</th>
<th>Tetragram</th>
<th>Frekvens</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>N_HA</td>
<td>0.0068</td>
</tr>
<tr>
<td>52</td>
<td>N OC</td>
<td>0.0068</td>
</tr>
<tr>
<td>53</td>
<td>AV</td>
<td>0.0066</td>
</tr>
<tr>
<td>54</td>
<td>ALL</td>
<td>0.0063</td>
</tr>
<tr>
<td>55</td>
<td>TER</td>
<td>0.0063</td>
</tr>
<tr>
<td>56</td>
<td>DE S</td>
<td>0.0063</td>
</tr>
<tr>
<td>57</td>
<td>ENNE</td>
<td>0.0061</td>
</tr>
<tr>
<td>58</td>
<td>ETT</td>
<td>0.0061</td>
</tr>
<tr>
<td>59</td>
<td>HEN</td>
<td>0.0060</td>
</tr>
<tr>
<td>60</td>
<td>T OC</td>
<td>0.0060</td>
</tr>
<tr>
<td>61</td>
<td>HENN</td>
<td>0.0060</td>
</tr>
<tr>
<td>62</td>
<td>E HA</td>
<td>0.0060</td>
</tr>
<tr>
<td>63</td>
<td>MEJ</td>
<td>0.0059</td>
</tr>
<tr>
<td>64</td>
<td>CH S</td>
<td>0.0059</td>
</tr>
<tr>
<td>65</td>
<td>MEJ</td>
<td>0.0059</td>
</tr>
<tr>
<td>66</td>
<td>T HA</td>
<td>0.0058</td>
</tr>
<tr>
<td>67</td>
<td>NNE</td>
<td>0.0057</td>
</tr>
<tr>
<td>68</td>
<td>TEN</td>
<td>0.0057</td>
</tr>
<tr>
<td>69</td>
<td>T VA</td>
<td>0.0057</td>
</tr>
<tr>
<td>70</td>
<td>NOM</td>
<td>0.0057</td>
</tr>
<tr>
<td>71</td>
<td>AN S</td>
<td>0.0056</td>
</tr>
<tr>
<td>72</td>
<td>ARA</td>
<td>0.0054</td>
</tr>
<tr>
<td>73</td>
<td>OM H</td>
<td>0.0054</td>
</tr>
<tr>
<td>74</td>
<td>N VA</td>
<td>0.0053</td>
</tr>
<tr>
<td>75</td>
<td>TT H</td>
<td>0.0053</td>
</tr>
<tr>
<td>76</td>
<td>EN D</td>
<td>0.0053</td>
</tr>
<tr>
<td>77</td>
<td>ETT</td>
<td>0.0053</td>
</tr>
<tr>
<td>78</td>
<td>SA</td>
<td>0.0053</td>
</tr>
<tr>
<td>79</td>
<td>EN H</td>
<td>0.0053</td>
</tr>
<tr>
<td>80</td>
<td>NAR</td>
<td>0.0052</td>
</tr>
<tr>
<td>81</td>
<td>N DE</td>
<td>0.0051</td>
</tr>
<tr>
<td>82</td>
<td>UPP</td>
<td>0.0051</td>
</tr>
<tr>
<td>83</td>
<td>ET V</td>
<td>0.0051</td>
</tr>
<tr>
<td>84</td>
<td>SKU</td>
<td>0.0050</td>
</tr>
<tr>
<td>85</td>
<td>VIL</td>
<td>0.0050</td>
</tr>
<tr>
<td>86</td>
<td>R HA</td>
<td>0.0050</td>
</tr>
<tr>
<td>87</td>
<td>KOM</td>
<td>0.0050</td>
</tr>
<tr>
<td>88</td>
<td>R AT</td>
<td>0.0050</td>
</tr>
<tr>
<td>89</td>
<td>UNDE</td>
<td>0.0050</td>
</tr>
<tr>
<td>90</td>
<td>FORS</td>
<td>0.0050</td>
</tr>
<tr>
<td>91</td>
<td>EN O</td>
<td>0.0049</td>
</tr>
<tr>
<td>92</td>
<td>SKUL</td>
<td>0.0049</td>
</tr>
<tr>
<td>93</td>
<td>KULL</td>
<td>0.0048</td>
</tr>
<tr>
<td>94</td>
<td>ET S</td>
<td>0.0048</td>
</tr>
<tr>
<td>95</td>
<td>VI</td>
<td>0.0048</td>
</tr>
<tr>
<td>96</td>
<td>SIG</td>
<td>0.0048</td>
</tr>
<tr>
<td>97</td>
<td>ULLE</td>
<td>0.0047</td>
</tr>
<tr>
<td>98</td>
<td>SIG</td>
<td>0.0047</td>
</tr>
<tr>
<td>99</td>
<td>T ME</td>
<td>0.0047</td>
</tr>
<tr>
<td>100</td>
<td>R OC</td>
<td>0.0047</td>
</tr>
</tbody>
</table>
B Språklig funktion

### B.1 Vinkelavstånd mellan bokstäver

Tabellen visar vinkelavstånd mellan bokstäver som togs fram i avsnitt 6.5.1. (Tabellen fortsätter på nästa sida).

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1.47</td>
<td>1.57</td>
<td>1.53</td>
<td>1.14</td>
<td>1.54</td>
<td>1.46</td>
<td>1.57</td>
<td>1.36</td>
<td>1.57</td>
<td>1.48</td>
<td>1.49</td>
<td>1.56</td>
<td>1.56</td>
<td>1.35</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1.47</td>
<td>0</td>
<td>1.56</td>
<td>1.06</td>
<td>1.54</td>
<td>1.06</td>
<td>1.36</td>
<td>1.05</td>
<td>1.56</td>
<td>1.07</td>
<td>1.09</td>
<td>1.38</td>
<td>1.10</td>
<td>1.51</td>
<td>1.54</td>
</tr>
<tr>
<td>C</td>
<td>1.57</td>
<td>1.56</td>
<td>0</td>
<td>1.56</td>
<td>1.57</td>
<td>1.56</td>
<td>1.57</td>
<td>1.56</td>
<td>1.56</td>
<td>1.56</td>
<td>1.56</td>
<td>1.57</td>
<td>1.57</td>
<td>1.57</td>
<td>1.57</td>
</tr>
<tr>
<td>D</td>
<td>1.53</td>
<td>1.06</td>
<td>1.56</td>
<td>0</td>
<td>1.55</td>
<td>1.46</td>
<td>1.21</td>
<td>1.36</td>
<td>1.56</td>
<td>1.44</td>
<td>1.40</td>
<td>1.40</td>
<td>1.03</td>
<td>1.27</td>
<td>1.56</td>
</tr>
<tr>
<td>E</td>
<td>1.14</td>
<td>1.54</td>
<td>1.57</td>
<td>1.55</td>
<td>0</td>
<td>1.56</td>
<td>1.54</td>
<td>1.57</td>
<td>1.33</td>
<td>1.57</td>
<td>1.53</td>
<td>1.52</td>
<td>1.57</td>
<td>1.56</td>
<td>1.48</td>
</tr>
<tr>
<td>F</td>
<td>1.54</td>
<td>1.06</td>
<td>1.56</td>
<td>1.46</td>
<td>1.56</td>
<td>0</td>
<td>1.39</td>
<td>1.30</td>
<td>1.56</td>
<td>1.31</td>
<td>1.27</td>
<td>1.46</td>
<td>1.40</td>
<td>1.52</td>
<td>1.55</td>
</tr>
<tr>
<td>G</td>
<td>1.46</td>
<td>1.36</td>
<td>1.57</td>
<td>1.21</td>
<td>1.54</td>
<td>1.39</td>
<td>0</td>
<td>1.47</td>
<td>1.55</td>
<td>1.47</td>
<td>1.34</td>
<td>1.46</td>
<td>1.34</td>
<td>1.08</td>
<td>1.56</td>
</tr>
<tr>
<td>H</td>
<td>1.57</td>
<td>1.05</td>
<td>1.56</td>
<td>1.36</td>
<td>1.57</td>
<td>1.30</td>
<td>1.47</td>
<td>0</td>
<td>1.57</td>
<td>0.85</td>
<td>0.95</td>
<td>1.47</td>
<td>1.28</td>
<td>1.53</td>
<td>1.57</td>
</tr>
<tr>
<td>I</td>
<td>1.36</td>
<td>1.56</td>
<td>1.57</td>
<td>1.56</td>
<td>1.33</td>
<td>1.56</td>
<td>1.55</td>
<td>1.57</td>
<td>0</td>
<td>1.57</td>
<td>1.51</td>
<td>1.54</td>
<td>1.57</td>
<td>1.56</td>
<td>1.52</td>
</tr>
<tr>
<td>J</td>
<td>1.57</td>
<td>1.07</td>
<td>1.56</td>
<td>1.44</td>
<td>1.57</td>
<td>1.31</td>
<td>1.47</td>
<td>0.85</td>
<td>1.57</td>
<td>0</td>
<td>1.08</td>
<td>1.43</td>
<td>1.33</td>
<td>1.34</td>
<td>1.57</td>
</tr>
<tr>
<td>K</td>
<td>1.48</td>
<td>1.09</td>
<td>1.56</td>
<td>1.40</td>
<td>1.53</td>
<td>1.27</td>
<td>1.34</td>
<td>0.95</td>
<td>1.51</td>
<td>1.08</td>
<td>0</td>
<td>1.38</td>
<td>1.39</td>
<td>1.43</td>
<td>1.55</td>
</tr>
<tr>
<td>L</td>
<td>1.49</td>
<td>1.38</td>
<td>1.56</td>
<td>1.40</td>
<td>1.52</td>
<td>1.46</td>
<td>1.46</td>
<td>1.47</td>
<td>1.54</td>
<td>1.43</td>
<td>1.38</td>
<td>0</td>
<td>1.41</td>
<td>1.43</td>
<td>1.56</td>
</tr>
<tr>
<td>M</td>
<td>1.56</td>
<td>1.10</td>
<td>1.56</td>
<td>1.03</td>
<td>1.57</td>
<td>1.40</td>
<td>1.34</td>
<td>1.28</td>
<td>1.57</td>
<td>1.33</td>
<td>1.39</td>
<td>1.41</td>
<td>0</td>
<td>1.31</td>
<td>1.57</td>
</tr>
<tr>
<td>N</td>
<td>1.56</td>
<td>1.51</td>
<td>1.57</td>
<td>1.27</td>
<td>1.56</td>
<td>1.52</td>
<td>1.08</td>
<td>1.53</td>
<td>1.56</td>
<td>1.34</td>
<td>1.43</td>
<td>1.43</td>
<td>1.31</td>
<td>0</td>
<td>1.56</td>
</tr>
<tr>
<td>O</td>
<td>1.35</td>
<td>1.54</td>
<td>1.57</td>
<td>1.56</td>
<td>1.48</td>
<td>1.55</td>
<td>1.56</td>
<td>1.57</td>
<td>1.52</td>
<td>1.57</td>
<td>1.55</td>
<td>1.56</td>
<td>1.57</td>
<td>1.56</td>
<td>1.56</td>
</tr>
<tr>
<td>P</td>
<td>1.54</td>
<td>1.28</td>
<td>1.57</td>
<td>1.41</td>
<td>1.56</td>
<td>1.33</td>
<td>1.38</td>
<td>1.39</td>
<td>1.56</td>
<td>1.40</td>
<td>1.39</td>
<td>1.42</td>
<td>1.37</td>
<td>1.47</td>
<td>1.56</td>
</tr>
<tr>
<td>Q</td>
<td>1.57</td>
<td>1.54</td>
<td>1.57</td>
<td>1.54</td>
<td>1.57</td>
<td>1.55</td>
<td>1.55</td>
<td>1.55</td>
<td>1.36</td>
<td>1.51</td>
<td>1.48</td>
<td>1.55</td>
<td>1.56</td>
<td>1.56</td>
<td>1.57</td>
</tr>
<tr>
<td>R</td>
<td>1.57</td>
<td>1.49</td>
<td>1.56</td>
<td>1.26</td>
<td>1.57</td>
<td>1.51</td>
<td>1.09</td>
<td>1.52</td>
<td>1.56</td>
<td>1.39</td>
<td>1.45</td>
<td>1.39</td>
<td>1.34</td>
<td>0.68</td>
<td>1.57</td>
</tr>
<tr>
<td>S</td>
<td>1.30</td>
<td>1.08</td>
<td>1.54</td>
<td>1.28</td>
<td>1.46</td>
<td>1.29</td>
<td>1.23</td>
<td>1.18</td>
<td>1.50</td>
<td>1.22</td>
<td>1.12</td>
<td>1.32</td>
<td>1.23</td>
<td>1.30</td>
<td>1.54</td>
</tr>
<tr>
<td>T</td>
<td>1.38</td>
<td>1.44</td>
<td>1.57</td>
<td>1.31</td>
<td>1.37</td>
<td>1.44</td>
<td>1.31</td>
<td>1.49</td>
<td>1.55</td>
<td>1.33</td>
<td>1.29</td>
<td>1.39</td>
<td>1.39</td>
<td>0.91</td>
<td>1.56</td>
</tr>
<tr>
<td>U</td>
<td>1.10</td>
<td>1.54</td>
<td>1.57</td>
<td>1.54</td>
<td>1.08</td>
<td>1.54</td>
<td>1.48</td>
<td>1.57</td>
<td>1.32</td>
<td>1.57</td>
<td>1.52</td>
<td>1.55</td>
<td>1.56</td>
<td>1.56</td>
<td>1.49</td>
</tr>
<tr>
<td>V</td>
<td>1.56</td>
<td>1.03</td>
<td>1.56</td>
<td>1.34</td>
<td>1.56</td>
<td>1.27</td>
<td>1.26</td>
<td>0.95</td>
<td>1.57</td>
<td>0.69</td>
<td>1.12</td>
<td>1.29</td>
<td>1.21</td>
<td>1.35</td>
<td>1.57</td>
</tr>
<tr>
<td>W</td>
<td>1.57</td>
<td>1.17</td>
<td>1.56</td>
<td>1.27</td>
<td>1.57</td>
<td>1.34</td>
<td>1.42</td>
<td>1.34</td>
<td>1.56</td>
<td>1.38</td>
<td>1.32</td>
<td>1.10</td>
<td>1.47</td>
<td>1.56</td>
<td>1.56</td>
</tr>
<tr>
<td>X</td>
<td>1.57</td>
<td>1.55</td>
<td>1.56</td>
<td>1.34</td>
<td>1.57</td>
<td>1.52</td>
<td>1.31</td>
<td>1.57</td>
<td>1.56</td>
<td>1.51</td>
<td>1.46</td>
<td>1.39</td>
<td>1.50</td>
<td>1.21</td>
<td>1.57</td>
</tr>
<tr>
<td>Y</td>
<td>1.34</td>
<td>1.55</td>
<td>1.57</td>
<td>1.54</td>
<td>1.43</td>
<td>1.55</td>
<td>1.50</td>
<td>1.57</td>
<td>1.38</td>
<td>1.56</td>
<td>1.52</td>
<td>1.53</td>
<td>1.56</td>
<td>1.54</td>
<td>1.48</td>
</tr>
<tr>
<td>Z</td>
<td>1.41</td>
<td>1.03</td>
<td>1.56</td>
<td>1.12</td>
<td>1.35</td>
<td>1.40</td>
<td>1.41</td>
<td>1.14</td>
<td>1.55</td>
<td>1.14</td>
<td>1.26</td>
<td>1.44</td>
<td>1.17</td>
<td>1.47</td>
<td>1.56</td>
</tr>
<tr>
<td>A</td>
<td>1.31</td>
<td>1.55</td>
<td>1.57</td>
<td>1.56</td>
<td>1.31</td>
<td>1.55</td>
<td>1.56</td>
<td>1.57</td>
<td>1.45</td>
<td>1.57</td>
<td>1.53</td>
<td>1.55</td>
<td>1.57</td>
<td>1.57</td>
<td>1.53</td>
</tr>
<tr>
<td>B</td>
<td>1.23</td>
<td>1.34</td>
<td>1.57</td>
<td>1.52</td>
<td>1.34</td>
<td>1.34</td>
<td>1.48</td>
<td>1.57</td>
<td>1.23</td>
<td>1.57</td>
<td>1.44</td>
<td>1.55</td>
<td>1.57</td>
<td>1.57</td>
<td>1.44</td>
</tr>
<tr>
<td>C</td>
<td>1.46</td>
<td>1.56</td>
<td>1.57</td>
<td>1.57</td>
<td>1.52</td>
<td>1.56</td>
<td>1.57</td>
<td>1.57</td>
<td>1.53</td>
<td>1.57</td>
<td>1.55</td>
<td>1.56</td>
<td>1.57</td>
<td>1.57</td>
<td>1.50</td>
</tr>
<tr>
<td>D</td>
<td>1.44</td>
<td>1.53</td>
<td>1.55</td>
<td>1.50</td>
<td>1.47</td>
<td>1.55</td>
<td>1.47</td>
<td>1.57</td>
<td>1.43</td>
<td>1.54</td>
<td>1.46</td>
<td>1.45</td>
<td>1.53</td>
<td>1.47</td>
<td>1.50</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td>J</td>
<td>K</td>
<td>L</td>
<td>M</td>
<td>N</td>
<td>O</td>
</tr>
<tr>
<td>P</td>
<td>1,54</td>
<td>1,57</td>
<td>1,57</td>
<td>1,30</td>
<td>1,38</td>
<td>1,10</td>
<td>1,56</td>
<td>1,57</td>
<td>1,57</td>
<td>1,34</td>
<td>1,41</td>
<td>1,31</td>
<td>1,23</td>
<td>1,46</td>
<td>1,44</td>
</tr>
<tr>
<td>Q</td>
<td>1,28</td>
<td>1,54</td>
<td>1,49</td>
<td>1,08</td>
<td>1,44</td>
<td>1,54</td>
<td>1,03</td>
<td>1,17</td>
<td>1,55</td>
<td>1,55</td>
<td>1,03</td>
<td>1,55</td>
<td>1,34</td>
<td>1,56</td>
<td>1,53</td>
</tr>
<tr>
<td>R</td>
<td>1,57</td>
<td>1,57</td>
<td>1,56</td>
<td>1,54</td>
<td>1,57</td>
<td>1,57</td>
<td>1,56</td>
<td>1,56</td>
<td>1,57</td>
<td>1,56</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,55</td>
</tr>
<tr>
<td>S</td>
<td>1,41</td>
<td>1,54</td>
<td>1,26</td>
<td>1,28</td>
<td>1,31</td>
<td>1,54</td>
<td>1,34</td>
<td>1,27</td>
<td>1,34</td>
<td>1,54</td>
<td>1,12</td>
<td>1,56</td>
<td>1,52</td>
<td>1,57</td>
<td>1,50</td>
</tr>
<tr>
<td>T</td>
<td>1,56</td>
<td>1,57</td>
<td>1,57</td>
<td>1,46</td>
<td>1,37</td>
<td>1,08</td>
<td>1,56</td>
<td>1,57</td>
<td>1,57</td>
<td>1,43</td>
<td>1,35</td>
<td>1,31</td>
<td>1,34</td>
<td>1,52</td>
<td>1,47</td>
</tr>
<tr>
<td>U</td>
<td>1,33</td>
<td>1,55</td>
<td>1,51</td>
<td>1,29</td>
<td>1,44</td>
<td>1,54</td>
<td>1,27</td>
<td>1,34</td>
<td>1,52</td>
<td>1,55</td>
<td>1,40</td>
<td>1,55</td>
<td>1,34</td>
<td>1,56</td>
<td>1,55</td>
</tr>
<tr>
<td>V</td>
<td>1,38</td>
<td>1,55</td>
<td>1,09</td>
<td>1,23</td>
<td>1,31</td>
<td>1,48</td>
<td>1,26</td>
<td>1,42</td>
<td>1,31</td>
<td>1,50</td>
<td>1,41</td>
<td>1,56</td>
<td>1,48</td>
<td>1,57</td>
<td>1,47</td>
</tr>
<tr>
<td>W</td>
<td>1,39</td>
<td>1,55</td>
<td>1,52</td>
<td>1,18</td>
<td>1,49</td>
<td>1,57</td>
<td>0,95</td>
<td>1,34</td>
<td>1,57</td>
<td>1,57</td>
<td>1,14</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
</tr>
<tr>
<td>X</td>
<td>1,56</td>
<td>1,57</td>
<td>1,57</td>
<td>1,56</td>
<td>1,56</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,14</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
</tr>
<tr>
<td>Y</td>
<td>1,57</td>
<td>1,56</td>
<td>1,34</td>
<td>1,23</td>
<td>1,39</td>
<td>1,56</td>
<td>1,21</td>
<td>1,10</td>
<td>1,50</td>
<td>1,56</td>
<td>1,17</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,53</td>
</tr>
<tr>
<td>Z</td>
<td>1,47</td>
<td>1,56</td>
<td>0,68</td>
<td>1,30</td>
<td>0,91</td>
<td>1,56</td>
<td>1,35</td>
<td>1,47</td>
<td>1,21</td>
<td>1,54</td>
<td>1,47</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,46</td>
</tr>
<tr>
<td>Å</td>
<td>1,56</td>
<td>1,57</td>
<td>1,57</td>
<td>1,54</td>
<td>1,56</td>
<td>1,49</td>
<td>1,57</td>
<td>1,56</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,56</td>
<td>1,53</td>
<td>1,44</td>
<td>1,50</td>
</tr>
<tr>
<td>Ä</td>
<td>0</td>
<td>1,56</td>
<td>1,51</td>
<td>1,20</td>
<td>1,49</td>
<td>1,55</td>
<td>1,35</td>
<td>1,47</td>
<td>1,54</td>
<td>1,56</td>
<td>1,44</td>
<td>1,42</td>
<td>1,48</td>
<td>1,57</td>
<td>1,54</td>
</tr>
<tr>
<td>Ö</td>
<td>1,56</td>
<td>0</td>
<td>1,56</td>
<td>1,56</td>
<td>1,56</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,56</td>
<td>1,56</td>
<td>1,55</td>
<td>1,57</td>
<td>1,46</td>
<td></td>
</tr>
<tr>
<td>_</td>
<td>1,51</td>
<td>1,56</td>
<td>0</td>
<td>1,33</td>
<td>1,07</td>
<td>1,56</td>
<td>1,32</td>
<td>1,44</td>
<td>1,16</td>
<td>1,55</td>
<td>1,45</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,47</td>
</tr>
<tr>
<td>1,20</td>
<td>1,56</td>
<td>1,33</td>
<td>0</td>
<td>1,21</td>
<td>1,28</td>
<td>1,05</td>
<td>1,07</td>
<td>1,43</td>
<td>1,46</td>
<td>1,21</td>
<td>1,48</td>
<td>1,52</td>
<td>1,54</td>
<td>1,47</td>
<td></td>
</tr>
<tr>
<td>1,49</td>
<td>1,56</td>
<td>1,07</td>
<td>1,21</td>
<td>0</td>
<td>1,54</td>
<td>1,32</td>
<td>1,32</td>
<td>1,31</td>
<td>1,50</td>
<td>1,11</td>
<td>1,48</td>
<td>1,51</td>
<td>1,56</td>
<td>1,47</td>
<td></td>
</tr>
<tr>
<td>1,55</td>
<td>1,57</td>
<td>1,56</td>
<td>1,28</td>
<td>1,54</td>
<td>0</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,33</td>
<td>1,55</td>
<td>1,36</td>
<td>1,28</td>
<td>1,51</td>
<td>1,47</td>
<td></td>
</tr>
<tr>
<td>1,35</td>
<td>1,57</td>
<td>1,32</td>
<td>1,05</td>
<td>1,32</td>
<td>1,57</td>
<td>0</td>
<td>0,90</td>
<td>1,46</td>
<td>1,56</td>
<td>1,08</td>
<td>1,57</td>
<td>1,56</td>
<td>1,57</td>
<td>1,54</td>
<td></td>
</tr>
<tr>
<td>1,47</td>
<td>1,57</td>
<td>1,44</td>
<td>1,07</td>
<td>1,32</td>
<td>1,57</td>
<td>0,90</td>
<td>0</td>
<td>1,51</td>
<td>1,57</td>
<td>1,18</td>
<td>1,57</td>
<td>1,55</td>
<td>1,57</td>
<td>1,52</td>
<td></td>
</tr>
<tr>
<td>1,54</td>
<td>1,56</td>
<td>1,16</td>
<td>1,43</td>
<td>1,31</td>
<td>1,57</td>
<td>1,46</td>
<td>1,51</td>
<td>0</td>
<td>1,56</td>
<td>1,50</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,44</td>
<td></td>
</tr>
<tr>
<td>1,56</td>
<td>1,57</td>
<td>1,55</td>
<td>1,46</td>
<td>1,50</td>
<td>1,33</td>
<td>1,56</td>
<td>1,57</td>
<td>1,56</td>
<td>0</td>
<td>1,52</td>
<td>1,46</td>
<td>1,38</td>
<td>1,42</td>
<td>1,33</td>
<td></td>
</tr>
<tr>
<td>1,44</td>
<td>1,56</td>
<td>1,45</td>
<td>1,21</td>
<td>1,11</td>
<td>1,55</td>
<td>1,08</td>
<td>1,18</td>
<td>1,50</td>
<td>1,52</td>
<td>0</td>
<td>1,55</td>
<td>1,56</td>
<td>1,57</td>
<td>1,46</td>
<td></td>
</tr>
<tr>
<td>1,42</td>
<td>1,56</td>
<td>1,57</td>
<td>1,48</td>
<td>1,48</td>
<td>1,36</td>
<td>1,57</td>
<td>1,57</td>
<td>1,46</td>
<td>1,55</td>
<td>0</td>
<td>1,43</td>
<td>1,49</td>
<td>1,50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,48</td>
<td>1,55</td>
<td>1,57</td>
<td>1,52</td>
<td>1,51</td>
<td>1,28</td>
<td>1,56</td>
<td>1,55</td>
<td>1,57</td>
<td>1,38</td>
<td>1,56</td>
<td>1,43</td>
<td>0</td>
<td>1,43</td>
<td>1,44</td>
<td></td>
</tr>
<tr>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,54</td>
<td>1,56</td>
<td>1,51</td>
<td>1,57</td>
<td>1,57</td>
<td>1,57</td>
<td>1,42</td>
<td>1,57</td>
<td>1,49</td>
<td>1,43</td>
<td>0</td>
<td>1,53</td>
<td></td>
</tr>
<tr>
<td>1,54</td>
<td>1,46</td>
<td>1,47</td>
<td>1,47</td>
<td>1,47</td>
<td>1,54</td>
<td>1,52</td>
<td>1,44</td>
<td>1,33</td>
<td>1,46</td>
<td>1,50</td>
<td>1,44</td>
<td>1,53</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
B.2 Likhet mellan bokstäver

Tabellen visar likheten i fråga om språklig funktion mellan samtliga bokstäver. Tabellen är rangordnad så att den bokstav som mest liknar den aktuella bokstaven, återfinns överst i tabellen. Övriga bokstäver kommer sedan i fallande skala.

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | Å | Ä | Ö |
| A | H | V | Z | Ä | V | D | V | E | B | B | S | W | G | Y | F | K | G | B | Z | S | H | M | T | A | T | U | A | A |
| Y | J | K | N | Z | J | X | S | Y | S | Z | X | S | G | B | V | P | X | Y | Z | D | L | Ö | H | I | F | U | L |
| O | S | B | W | T | P | T | M | F | F | T | H | M | Ö | K | H | S | Z | D | Ä | K | T | S | E | M | Y | Y | E | Z |
| I | K | L | S | Y | W | K | F | Ä | M | T | R | N | J | I | H | L | M | T | G | _ | M | L | _ | S | W | S | _ | K |
| Z | W | H | X | G | B | D | K | N | L | M | R | L | S | D | G | L | M | V | O | F | H | R | K | K | O | K | O | S | T |
| K | T | F | F | P | G | L | L | R | G | G | D | N | F | P | T | N | T | B | N | L | F | N | N | F | L | P | Z | T | L | S |
| D | R | T | F | D | X | H | Q | B | X | R | F | X | F | P | T | P | P | F | L | F | L | X | X | B | D | R | L | D | P | D |
| P | N | G | P | N | C | D | Q | H | H | D | R | X | H | R | Z | C | B | G | L | Z | O |
| F | F | P | A | F | A | J | Y | F | X | A | U | Y | L | X | A | Y | X | A | P | C | A | Q | F | N | D | Q | Q | A |
| N | U | A | N | U | A | X | P | C | Q | E | C | A | Z | A | M | Q | Y | P | M | A | Y | X | Q | W | D | W |
| V | E | A | Q | V | L | R | Y | A | Y | Y | Q | N | O | C | E | H | N | Y | O | I | P | Y | V | Z | V | O |
| M | O | Y | U | Q | A | Y | W | I | I | Q | U | Q | U | A | L | Y | R | A | T | H | J | A | W | V | J | M |
| R | Q | I | Y | M | Q | E | O | N | A | Y | Q | A | E | V | E | W | I | A | A | W | E | Y | A | V | I | N | N | H | B |
| H | Y | U | E | R | O | I | I | X | O | U | A | O | O | M | Y | E | A | U | I | U | U | X | U | M | J | W | V |
| X | A | Q | I | W | Y | E | Q | E | E | E | U | E | I | J | O | Y | E | A | I | V | O | Q | E | M | Q | R | M | N | P |
| Q | X | E | Á | X | E | O | A | J | A | A | Å | Å | I | C | H | U | O | O | O | Q | U | O | O | O | Q | E | O | H | C | J | C | M | J |
| W | C | Å | C | H | I | C | U | H | Å | O | O | A | Å | X | C | O | Å | O | C | A | O | Q | O | X | X | R | C |
| C | O | Ö | O | C | C | Ö | A | C | O | C | O | Ö | O | C | Ö | C | O | Q | C | O | Q | C | H | Ö | O | A | C | Ö | C | Ö | C | H | C | H |
C Testtexter

C.1 Testtext A-E
Här redovisas testtext A-E som använts som jämförelseunderlag i kapitel 6. Texterna återges precis som de använts, dvs skiljetecken och övriga tecken som inte finns bland de 30 symboler vi betraktat genom arbetet, har uteslutits.

C.1.1 Testtext A
Testtext A är ett fritt översatt citat av Albert Einstein.

    Jag vet ej med vilka vapen det tredje världskriget kommer att
    utkämpas men det fjärde kommer att utkämpas med påkar och stenar

C.1.2 Testtext B
Testtext B utgör första versen av den kända svenska psalmen ”Den blomstertid nu kommer”

    Den blomstertid nu kommer Med lust och fägring stor Du nalkas ljusa
    sommar Då gräs och gröda gror Med blid och livlig värma Till allt
    som varit dött sig solens strålar närma Och allt blir återfött

C.1.3 Testtext C
Testtext C är ett stycke ur den svenska tryckfrihetsförordningen.

    Med tryckfrihet förstås varje svensk medborgares rätt att utan några
    av myndighet eller annat allmänt organ i förväg lagda hinder utgiva
    skrifter att sedermera endast inför laglig domstol kunna tilltalas för
    deras innehåll och att icke i annat fall kunna straffas därför än om
    detta innehåll strider mot tydlig lag given att bevara allmänt lugn utan
    att återhålla allmän upplysning I överensstämmelse med de i första
    stycket angivna grunderna för en allmän tryckfrihet och till
    säkerställande av ett fritt meningsutbyte och en allsidig upplysning
    skall det stå varje svensk medborgare fritt att med iakttagande av de
    bestämmelser som äro i denna förordning meddelade till skydda för
    enskild rätt och allmän säkerhet i tryckt skrifte yttra sina tankar och
    äsikter offentliggöra allmänna handlingar samt meddela uppgifter och
    underrättelser i vad ämne som helst. Det skall så stå envan fritt att i
    alla de fäll då ej annat är i denna förordning föreskrivet meddela
    uppgifter och underrättelser i vad ämne som helst för offentliggörande
    i tryckt skrifte till författare eller annan som är att anse som
    upphovsman till framställning i skriften till skriften utgivare eller om
Förenklad textinmatning på mobila enheter
med hjälp av kontextbaserad språktolkning

för skriften finnes särskild redaktion till denna eller till företag för
yrkesmässig förmedling av nyheter eller andra meddelanden till
periodiska skrifter Vidare skall envar äga rätt att om ej annat följer av
denna förordning anskaffa uppgifter och underrättelser i vad ämne
som helst för att offentliggöra dem i tryckt skrift eller för att lämna
meddelande som avses i föregående stycke

C.1.4Testtext D

Testtext D består av inledningen till August Strindbergers roman
"Hemsöborna".

Han kom som ett yrväder en aprilafton och hade ett höganäskrus i en
svångrem om halsen Clara och Lotten var inne med skötekan att
hämta honom på Dalarö brygga men det dröjde evigheter innan de
kam in båt De skulle till handelsman och ha en tunna tjära och på
abeteket och hämta gråsalva åt grisen och så skulle de på posten och
få ett frimärke och så skulle de ner till Fia Lövström i Kroken och låna
tuppen mot ett halvpund småtärna till notbygget och sist hade de
hamnat på gästgivaregården där Carlsson bjudit på kaffe med dopp

C.1.5 Testtext E

Testtext E är ett autentiskt textmeddelande.

tjena Jag kommer nog inte hem förr än sent inatt eftersom jag måste
jobba övert Vänta inte uppe på mig.
På svenska

Detta dokument hålls tillgängligt på Internet – eller dess framtida ersättare – under en längre tid från publiceringsdatum under förutsättning att inga extraordinära omständigheter uppstår.

Tillgång till dokumentet innebär tillstånd för var och en att läsa, ladda ner, skriva ut enstaka kopiör för enskilt bruk och att använda det oförändrat för ickekommersiell forskning och för undervisning. Överföring av upphovsrätten vid en senare tidpunkt kan inte upphäva detta tillstånd. All annan användning av dokumentet kräver upphovsmannens medgivande. För att garantera äktheten, säkerheten och tillgängligheten finns det lösningar av teknisk och administrativ art.

Upphovsmannens ideella rätt innefattar rätt att bli nämnd som upphovsman i den omfattning som god sed kräver vid användning av dokumentet på ovan beskrivna sätt samt skydd mot att dokumentet ändras eller presenteras i sådan form eller i sådant sammanhang som är kränkande för upphovsmannens litterära eller konstnärliga anseende eller egenart.

För ytterligare information om Linköping University Electronic Press se förlagets hemsida http://www.ep.liu.se/

In English

The publishers will keep this document online on the Internet - or its possible replacement - for a considerable time from the date of publication barring exceptional circumstances.

The online availability of the document implies a permanent permission for anyone to read, to download, to print out single copies for your own use and to use it unchanged for any non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional on the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility.

According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement.

For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its WWW home page: http://www.ep.liu.se/

© Anders Jensen