Originalspråk och Översättningssvenska: En Korpusbaserad Jämförelse

Fördjupningsuppgift i 729G49

1 Introduktion

Målet med fördjupningsuppgiften är att genomföra en korpusstudie som kombinerar kvantitativa metoder (med hjälp av programmeringstekniker från språkteknologidelen) och kvalitativa analyser (baserade på era lingvistiska kunskaper).

2 Forskningsfrågor

Vilka spår lämnar översättningsprocessen i maskinöversatt svenska, och hur kan de identifieras med korpuslingvistiska metoder?

- Leder maskinöversättning till lägre lexikal variation i svensk text?
- Vilka ord och konstruktioner försvinner, och vilka överrepresenteras?

3 Data

I det här projektet jämför du två svenska textkorpusar med hjälp av beräkningslingvistiska metoder:

- **Originalkorpusen** innehåller artiklar från Göteborgs-Posten (SpråkbankenText, 2026), dvs. autentisk, journalistisk svenska.
- **Backtranslationskorpusen** innehåller samma meningar, men automatiskt översatta till engelska och sedan tillbaka till svenska med maskinöversättningsmodellen OPUS-MT (Tiedemann et al., 2023; Tiedemann and Thottingal, 2020).

Båda korpusarna inkluderar lingvistisk annotation, precis som i labbarna: de har parsats med spaCy (Honnibal et al., 2020). Obs att den automatiska taggningen är långt ifrån perfekt! Filarna är i CoNLL-X-format, samma format som i labbarna.

4 Uppgift

Uppgiften är att skriva ett Python-program som läser in de två filerna och genomför analyserna nedan. Du ska både redovisa resultaten och tolka dem: Vad säger skillnaderna om hur maskinöversatt språk skiljer sig från originalspråk?

4.1 Steg 1: Grundläggande korpusstatistik

Beräkna följande mätvärden för vardera korpus och presentera dem i en tabell:

- Antal meningar
- Antal token (totalt antal ord)
- Antal typer (unika ordformer)
- Antal typer baserat på lemma
- TTR (Type-Token Ratio)
- TTR på bigramnivå

Diskutera: Vilken korpus har högre TTR och bigram-TTR, och vad säger det om maskinöversättningens effekt på lexikal variation?

4.2 Steg 2: Överrepresenterade ord

Hitta ord som förekommer proportionellt mycket oftare i den ena korpusen än i den andra.

- Räkna hur många gånger varje lemma förekommer i respektive korpus. Exkludera token med ordklasser som inte är relevanta (t.ex. interpunktionstecken och siffror).
- Beräkna den relativa frekvensen för varje lemma: antal förekomster dividerat med totalt antal token i korpusen.
- Beräkna kvoten för varje lemma: relativ frekvens i originalet dividerat med relativ frekvens i backtranslationen.
- Inkludera bara lemman som förekommer minst 10 gånger i *båda* korpusarna.
- Analysera de 50 lemman med högst kvot (överrepresenterade i originalet) och de 50 med lägst kvot (överrepresenterade i backtranslationen).

Diskutera: Vilka typer av ord är överrepresenterade i originalet respektive backtranslationen? Ser

du mönster kopplade till stil (t.ex. talspråklighet, formalitet) eller till vanliga drag inom översättningssvenska?

4.3 Steg 3: Överrepresenterade kollokationer

Tillämpa samma frekvenskvot-metod som i Steg 2, men på ordpar (bigram av lemman) i stället för enskilda ord.

- Extrahera alla på varandra följande lemmapar inom varje mening. Exkludera även här token med icke-relevanta ordklasser.
- Räkna bigramfrekvenser i varje korpus och beräkna relativa frekvenser.
- Beräkna kvoten och filtrera bort bigram med färre än 10 förekomster i endera korpusen.
- Presentera de mest överrepresenterade bigrammen i varje riktning.

Diskutera: Avslöjar de överrepresenterade bigrammen typiska fraser eller konstruktioner som skiljer originalet från backtranslationen?

4.4 Steg 4: Kvalitativ analys av kollokationer

Välj tre bigram från Steg 3 som du tycker är intressanta och analysera dem närmare. För varje bigram:

- Hitta exempel i backtranslationskorpusen där bigrammet förekommer.
- Hitta motsvarande mening i originalkorpusen och i den engelska mellanöversättningen.
- Jämför de tre versionerna och diskutera: Vad är det i den engelska översättningen som ger upphov till bigrammet i backtranslationen? Vilken formulering används i originalet, och vad händer med den i översättningsprocessen?

Diskutera: Vad avslöjar dina tre exempel om hur översättningsprocessen förändrar svenska formuleringar? Koppla gärna till mönster du har sett i de kvantitativa analyserna.

4.5 Steg 5: Sammanfattande diskussion

Sätt de olika delarna i sammanhang. Vilka tecken på översättningssvenska kunde analyserna avslöja? Vad saknas, och vilka metoder skulle behövas för att hitta det?

5 Del 2 (välj 1 för VG)

Del 1 var utformad för att ge ganska tydliga resultat. Uppgifterna i Del 2 är mer explorativa: du ska i högre grad designa analysen själv, och resultaten kommer inte nödvändigtvis vara lika lätta att tolka. Välj **ett** av alternativen nedan.

5.1 Alternativ 1: Inkludera LLM-översättningar

Det finns indikationer på att stora språkmodeller producerar mindre ordagranna översättningar än traditionella maskinöversättningssystem (Raunak et al., 2023; Kunz et al., 2026). Vi har därför översatt GP-korpusen även med Gemma-3-12B-it (GemmaTeam, 2025), en stor språkmodell tränad av Google. Inkludera denna tredje korpus i din studie. Tillämpa samma metoder som i Del 1 och jämför de tre korpusarna och undersök:

- Liknar LLM-översättningen mer OPUS-MT-versionen eller originaltexterna? Varierar svaret beroende på vilken metod du använder?
- Vilka konkreta skillnader finns mellan de två översättningssystemen? Kan du identifiera ord eller konstruktioner som är karaktäristiska för det ena men inte det andra?

Diskutera dina resultat i relation till relevant litteratur om hur LLM-baserad översättning skiljer sig från neurala maskinöversättningssystem.

5.2 Alternativ 2: Domäneffekter

I Del 1 undersökte du journalistisk text från Göteborgs-Posten. Men påverkar maskinöversättning alla typer av text på samma sätt? Vi tillhandahåller en ytterligare korpus bestående av utdrag ur Selma Lagerlöfs verk, filtrerad från en korpus av äldre svenska romaner (SpråkbankenText, 2017). Korpusen har backtranslaterats med samma metod som GP-korpusen (svenska → engelska → svenska via OPUS-MT). Tillämpa metoderna från Del 1 på Lagerlöf-korpusen och jämför med dina resultat från GP-korpusen:

- Är maskinöversättningens effekter på TTR och lexikal variation större eller mindre för litterär text än för tidningstext?
- Vilka typer av ord och bigram är överrepresenterade i originalet respektive backtranslationen? Skiljer sig mönstren från dem du såg i GP-korpusen?
- Hanterar maskinöversättningen ålderdomligt eller litterärt språk på ett annat sätt än modern svenska? Moderniseras texten, förenklas den, eller förändras den på andra sätt?

Diskutera dina resultat i relation till relevant litteratur om hur domän och textstil påverkar maskinöversättningskvalitet.

6 Rapport

Introduktion Ge en översikt över problemområdet och beskriv syftet med din studie. Avsluta med 1–3 tydligt formulerade **forskningsfrågor** som du besvarar i rapporten, antingen från Kapitel 2 eller dina egna. Det ska redan i introduktionen framgå vad du har undersökt och hur.

Litteraturöversikt Diskutera minst tre relevanta vetenskapliga artiklar och hur de relaterar till din studie.

Metodik Beskriv i detalj hur du har genomfört studien. Fokus ska ligga på metoden: datan, hur du har valt ut ord och konstruktioner, och hur du har implementerat analysen. Inkludera dock **ingen Pythonkod**!

Resultat och diskussion Presentera de **kvantitativa** resultaten på ett neutralt och tydligt sätt, till exempel i tabellform. Den **kvalitativa** diskussionen ska innehålla konkreta exempel från korpusen, analyserade med lingvistiska begrepp, och relatera dina resultat till tidigare forskning från litteraturöversikten.

Begränsningar Gör en metodkritik. Vilka slutsatser kan du **inte** dra utifrån studien? Vad skulle kunna förbättras – till exempel val av korpus, analysmetod, kod eller urval av exempel?

Slutsatser Sammanfatta de viktigaste resultaten. Besvara forskningsfrågan tydligt och kortfattat. Lyft gärna fram förslag på vidare studier.

7 Betygskriterier (G)

Vi bedömer rapporterna främst efter följande kriterier:

- Rapporten ger en tydlig översikt av problemområdet och studien och utgår från en relevant forskningsfråga. Strukturen i Avsnitt 6 följs.
- Relevant litteratur diskuteras i relation till den egna studien på ett meningsfullt sätt.
- Lingvistiska begrepp används korrekt och integreras konsekvent i analysen.
- Koden är korrekt och begriplig och producerar alla resultat som beskrivs i rapporten.
- Slutsatserna baseras på både kvantitativa och kvalitativa resultat. Du visar förståelse för studiens begränsningar.

8 Betygskriterier (VG)

För att uppnå betyget VG ska du bygga ut studien och forskningsfrågorna med en av VG-alternativen och ta upp den utökade analysen på ett meningsfullt sätt i diskussionen. Du ska också ha:

- En utförlig diskussion av relevant litteratur, inklusive minst tre vetenskapliga texter som du själv har hittat genom egen litteratursökning (som inte finns med i det här dokumentet).
- En kreativ utökning av studien, antingen med minst ett till relevant mått eller med en mer omfattande kvalitativ analys.
- Högkvalitativ, läsbar kod som visar på språkteknologisk och lingvistisk förståelse.
- En omfattande diskussion av studiens begränsningar, inklusive konkreta förslag på hur några av dem kan adresseras.

References

GemmaTeam. 2025. Gemma 3.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Jenny Kunz, Anja Jarochenko, and Marcel Bollmann. 2026. A dataset for probing translationese preferences in english-to-swedish translation. *Preprint*, arXiv:2603.08450.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. Association for Computational Linguistics.

SpråkbankenText. 2017. Äldre svenska romaner.

SpråkbankenText. 2026. Göteborgsposten.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.