# Framtiden för lingvistisk validering är här:
## Ökad efterlevnad av konsekvens-gran skningsprocesser med GenAI

En presentation från Lionbridge med Elisabet Sas Olesa, Karolina Elizondo, Kathryn Nolte, Nathalie Azuaje och Melinda Johnson

## INLEDNING

**I studien undersökte vi hur generativ AI (GenAI) kan vara till hjälp vid lingvistisk validering (LV), framför allt med fokus på hur tekniken kan effektivisera konsekvensgranskningssteget vid dubbel översättning.** Dubbel översättning innebär att två eller flera oberoende översättningar förenas till en (Koller et al., 2012).

Det huvudsakliga målet för vår undersökning var att utvärdera GenAI:s förmåga att upptäcka avvikelser i de konsekvensgranskningar som görs med konventionella metoder.

## METODER

**Vi genomförde en praktisk analys av ett PerfO-stickprov (1 000–2 000 ord) som innehöll en blandning av språk från Asien-Stillahavsregionen, long tail-språk och regionala språkvarianter.** Utifrån de här parametrarna valde vi ut olika filer från en tidigare genomförd lokalisering av Clinical Evaluation of Language Fundamentals®-Fifth Edition (CELF-5), ett flexibelt system med tester som ges individuellt. Testerna används ofta för att underlätta diagnostisering av språksvårigheter hos barn och ungdomar (NCS Pearson, Inc., 2013). De målspråk som valdes ut för analys var spanska (Argentina), spanska (Spanien), franska (Frankrike), armeniska (Armenien), japanska (Japan) samt traditionell kinesiska (Taiwan).

## UTVECKLAD PROMPTDESIGN

**Den första prompten utformades för att arbeta parallellt med det traditionella resultatet av konsekvensgranskningen genom att jämföra översättning A mot översättning B. På så sätt skapades ett binärt resultat med godkända och underkända förslag utifrån granskarens slutgiltiga beslut.**

Det visade sig dock att den här metoden hade avsevärda begränsningar, i synnerhet när den används till andra språk än engelska och i väldigt specialiserade sammanhang – till exempel de som gäller inom specifika behandlingsområden eller för sjukdomar/åkommor som vanligtvis bedöms med livskvalitetsinstrument. Som en reaktion på rönen valde vi att använda GenAI för att utvärdera konsekvensgranskningen och därmed understödja vårt arbetsflöde för kvalitetssäkring.

Den uppdaterade prompten integrerades sedan i *Aurora Clinical Outcomes*, Lionbridges egen plattform för heltäckande lingvistisk validering. Genom att använda GenAI som stöd i den interna beslutsprocessen fick vi även insikter om konsekvensgranskarens tankegångar. Därmed säkerställde vi att uppgiften utfördes enligt nödvändiga standarder och branschkrav. Kraven definieras i MÅL FÖR KONSEKVENSGRANSKNING och lyfts fram i den vägledning som ges till granskaren i arbetsinstruktionerna:

1. **Begreppsmässig överens-stämmelse med det ursprungliga måttet**
2. **Kulturell anpassning**
3. **Tillgänglighet för studiens avsedda befolkningsgrupp/målgrupp**
4. **Detektering av fördomstendenser**

## RESULTAT

### Positiva resultat

När GenAI-resultatet analyserades av språkexperter identifierades tre viktiga användningsområden:

- **Felaktiga eller ofullständiga skäl:** GenAI identifierade genomgående kommentarer som saknade tillräckligt starka språkliga eller begreppsmässiga argument (t.ex. ”Översättning A är bättre” eller ”Jag föredrar översättning B”). Den upptäckte även olösta problem och obesvarade frågor. Det ledde fram till rekommendationer om att finslipa prompten för att se till att sådana förekomster eskaleras och följs upp.
- **Avsaknad av skäl:** Samtliga experter noterade att GenAI tillförlitligt kunde identifiera avsaknad av, ofullständiga och otydliga skäl, vilket ledde till betydligt snabbare kvalitetssäkring.
- **Effektivitet:** GenAI har mycket hög arbetstakt och kan analysera upp till 300 textsegment på bara några sekunder.

Med hjälp av den här gallringen kan språkexperter snabbare filtrera resultaten och avgöra om en fil är redo att skickas vidare till nästa steg i den lingvistiska valideringen.

### Förbättringsområden

Även om GenAI har stor potential är den också oförutsebar till sin natur och uppvisade vissa begränsningar när prompten kördes:

- **Inkonsekvent beteende:** GenAI gav inkonsekventa resultat vid utvärdering av identiska kommentarer i olika segment, till och med i samma fil. Det visar att GenAI-resultaten inte är helt tillförlitliga som fristående lösning.
- **Svårigheter kopplade till sammanhang/metareferenser:** Återkommande användning av vaga kommentarer och referenser som ”Se kommentaren ovan” skapar stora svårigheter för GenAI i den nuvarande versionen. GenAI kan inte hitta den tidigare kommentar som hänvisningen avser. Experterna föreslog att granskaren ska använda konkreta och tydliga referenser och att förklaringar ska upprepas i varje segment.
- **Vanliga kommentarer i konsekvensgranskningar är inte standardiserade:** Våra experter är överens om att alla inblandade behöver komma överens om mer exakta och tydliga definitioner av godtagbara resonemang och beskrivningar av skäl.

## SLUTSATS

**Studien visar att det finns stor potential i att använda GenAI som kvalitetsgranskningsverktyg för att upptäcka luckor, icke-efterlevnad och utelämnade skäl.**

GenAI kan användas för att snabbare upptäcka innehåll som behöver arbetas om, så att lokaliseringsteamet kan ge mer konkret feedback till intressenter med insikter om hur de kan förbättra sina resultat. Mänsklig expertis är emellertid fortfarande avgörande för att tolka resultat, validera detaljerade beslut och säkerställa överensstämmelse med projektkriterier och regulatoriska krav. En metod med ”en människa i processen” (som kombinerar GenAI-baserad analys med granskning av experter) höjer den övergripande kvaliteten, förbättrar granskarens resultat genom insiktsfull feedback och bidrar till att processen kontinuerligt finslipas.

Genom att kontinuerligt optimera promptar och ge motorn rätt träning kan metoden i slutändan leda till resultat av högre kvalitet och ökad kostnadseffektivitet i COA-lokaliseringsprojekt.

Citat:
Koller, M. et al. (2012). The process of reconciliation: evaluation of guidelines for translating quality-of-life questionnaires. Expert Review of Pharmacoeconomics and Outcomes Research 12(2), s. 189–197.

# GenAI i praktiken:
## En studie i effektiviteten hos AI-baserad jämförande granskning

Presenterad av Lionbridge: Stephanie Casale

## MÅLSÄTTNINGAR

**Lingvistisk validering (LV) är en process för att lokalisera och granska kliniska utfallsbedömningar för att säkerställa att de data som samlas in på olika platser är korrekta och konsekventa.** Processen är både omfattande och komplicerad till sin natur. Metoden säkerställer att översättningarna blir så noggranna som möjligt och håller högsta kvalitet, men komplexiteten har ett pris. I ett försök att minska både kostnaden och arbetstiden i processen har vår studie som målsättning att korta ledtider och outsourcingkostnader och samtidigt bibehålla den höga standard som processen kräver.

Den här affischen undersöker möjligheten att använda generativ AI för att utföra jämförande granskning. Jämförande granskning är ett viktigt steg i kvalitetssäkringen under lingvistisk validering. Den innebär att källtext och återöversättning jämförs för att säkerställa begreppsmässig överensstämmelse. Steget är ett led i en längre process där föregående och senare steg utförs av erfarna och utbildade språkexperter. Därför lämpar sig jämförande granskning väl för automatisering. Metoden minimerar risken för oupptäckta fel innan innehållet färdigställs.

Målsättningen för vår studie var att utforma en prompt som levererar minst lika hög kvalitet som de språkexperter som i dag utför jämförande granskning.

## METODER

**Vi började med att utforma en prompt som gav förväntat resultat från en jämförande granskning och en kommentar från en jämförande granskning med ytterligare information om resultatet. Resultatet av den jämförande granskningen delades in i tre kategorier:**

**Identiska** – Det här resultatet visar att källtexten och återöversättningen var exakt likadana på alla sätt, även när det gäller versaler/gemener och skiljetecken.

**Motsvarande** – Resultatet visar att även om det förekom skillnader i ordval, meningsbyggnad eller andra detaljer stämde segmenten överens rent begreppsmässigt. Läsaren skulle förstå att de förmedlar samma information.

**Behöver granskas** – Resultatet visar att det finns något i de två segmenten som gör att de inte stämmer överens begreppsmässigt. En läsare skulle kunna missförstå den översatta texten och tro att den betydde något som inte avsetts med källtexten.

Därefter utformades en prompt för att producera en kommentar om en jämförande granskning för alla icke-identiska resultat. Kommentarerna förklarar begreppsmässiga skillnader mellan de två segmenten, med beskrivningar av hur en lekman kan missuppfatta dem. Prompten ombads ignorera eventuella skillnader i skiljetecken och versaler/gemener, såvida de inte var direkt relaterade till betydelse och förståelse. Prompten skulle även ignorera eventuell extra text som inte var relaterad till källtextens betydelse (t.ex. formateringsmärkning).

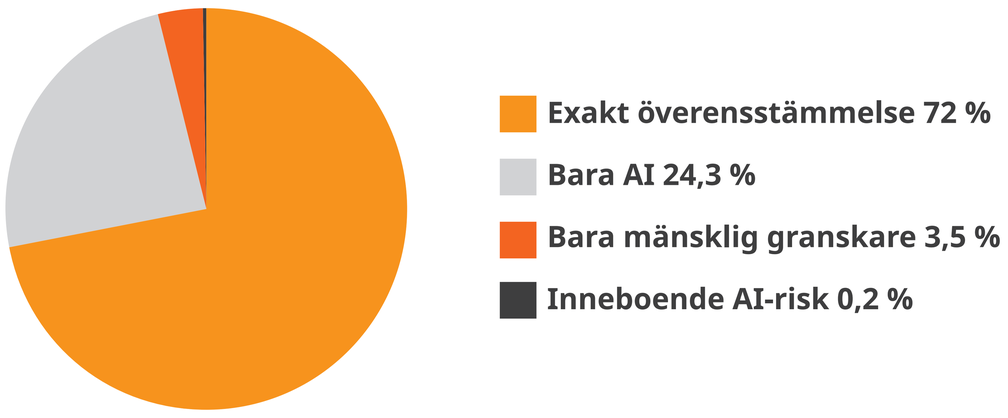
För att sammanställa data använde vi en tidigare lokaliserad bedömning från Pearson, Delis-Kaplan Executive Function System (D-KEFS) - Trail Making Test. Testet innehåller ca 1 000 källord. Återöversättningarna utfördes av olika språkexperter som hade antingen engelska eller målspråket som modersmål. Både en professionell jämförande granskare och medarbetare i vårt eget COA-lokaliseringsteam utförde flera jämförande granskningar. Erfarenheten inom jämförande granskning skilde sig åt i gruppen, som hade olika modersmål. På så sätt kunde vi få fram varierade resultat för jämförelse.

## RESULTAT

**De första resultaten är lovande, med tydliga och konkreta beskrivningar av den ursprungliga granskningen och avvikelser i återöversättningen med en övergripande preliminär korrekthet på 96,4 %.**

Uppdelade resultat presenteras i följande diagram. Det visar bland annat 72,09 % exakt överensstämmelse när manuell granskning jämfördes med AI-granskning. Ytterligare 24,3 % av segmenten flaggades som avvikelser av AI, men inte av den mänskliga granskaren. Resultaten innehöll 3,5 % avvikelser som flaggats av den mänskliga granskaren, men som AI ansåg stämde överens. Resultaten omfattade även 0,17 % identiska svar, vilket forskarna betraktade som riskabelt, eftersom översättningen inte var skriven på det latinska alfabetet och därmed lätt kunde missas när hänsyn inte togs till detta.

### ANDEL AI-RESULTAT JÄMFÖRT MED MANUELLA RESULTAT




**Intressanta procenttal:**
De GenAI-resultat som analyserades av våra språkexperter visade att det finns tre huvudsakliga områden där GenAI kan erbjuda effektivt stöd:

**Inneboende AI-risk – 0,17 %:** I den här siffran ingår segment som eventuellt flaggades av en människa på grund av att översättningen inte var skriven på det latinska alfabetet och detta hade inte upptäckts av AI-prompten.

**Inkonsekventa svar – 1,26 %:** Under studien hände det att AI gav olika svar för samma uppsättning segment, vilket utgjorde något över 1 % av den totala datamängden.

## SLUTSATS

**Generativ AI skulle kunna ge betydande tids- och kostnadsbesparingar vid lingvistisk validering.**

Det behövs ytterligare studier med en större datauppsättning. I kommande steg bör det ingå ett koncepttest för att undersöka effekterna av att använda resultaten tillsammans med språkexperter vid jämförande granskning under konsekvensgranskning. Prompten behöver finkalibreras ytterligare för att förhoppningsvis minska vissa inkonsekvenser och risker.