Projektnummer
A12512
Uppdragsgivare
Försvarsdepartementet
FoT-område
Inget FoT-område

Författare
Selbi Gustafsson, Sofiya Voytiv

Datum
2025-12-16
Memo nummer
FOI Memo 9166

Utvärdering av datakvalitén i The BB project

1 Inledning

Syftet med detta memo är att utvärdera användningen av *tal-till-text* transkriberingsplattformen BB-project för att kunna utföra omvärldsanalys. BB-project arkiverar kontinuerligt sändningar från 23 ryskspråkiga tv-kanaler från 2020-01-01 och framåt, och laddar upp dem med en timmes fördröjning på sin webbplats. På webbplatsen kan man se både video och automatiska *tal-till-text* transkriberingar. Webbplatsen saknar information om hur transkriberingen genomförs, alltså vilken ASR (Automatic Speech Recognition)-pipeline (automatiserade transkriptionsmodeller) som används eller om dess kvalité.

Listan över kanaler inkluderar både propagandistiska, statsfinansierade kanaler som RBK TV och Rossiya-24, och kanaler som inte är kända för desinformation, såsom Euronews Russia och Dozhd. Dozhd har även klassats i Ryssland som ”utländsk agent” sedan 2021. BB-project ger dessutom tillgång till den belarussiska statsfinansierade kanalen Belarus 24 och den kinesiska statliga kanalen (CGTN) som sänds på ryska. Kanalutbudet omfattar både nyheter och politiska program men också underhållning såsom serier och filmer.

Memot börjar med en beskrivning av vårt urval av datamaterial för utvärdering. Sedan diskuterar vi studiemetoden. Därefter redovisar vi resultatet och avslutar memot med rekommendationer för användning av data från BB-project i framtida studier.

2 Data och urval

Analysen gjordes på två kanaler som bedöms vara mest relevanta för FOI:s verksamhet. Mer specifikt är de finansierade av den ryska staten och kända för en hög grad av desinformation och/eller propaganda – RBK TV och Rossiya-24 (Daukšas m.fl., 2024). Både RBK TV och Rossiya-24 har huvudsakligen politiskt innehåll som nyheter och debatt, där RBK TV fokuserar mest på nyheter medan Rossiya-24 även sänder olika politiska program. Samtliga kanaler innehåller reklam. En slumpmässig sändningstid valdes: 15 juni 2025 mellan kl. 10:00 och 11:00 för båda kanalerna. Detta gav två timmars sändningsmaterial. Transkriberingar från båda kanalerna laddades ned, där RBK TV innehöll 6 598 ord och Rossiya-24 – 5 427 ord. För analysen använde vi all text som transkriberats inom tidsramen inklusive reklam, nyheter och politiska program och dokumentärer. Syftet var att utvärdera språket och därför var programinnehållet mindre avgörande. Samtidigt var det relevant att bedöma hur väl BB:s algoritmer hanterar specifika ryska termer som ofta förekommer i krigsrapporteringar. Därför valdes inte kanaler som främst sänder underhållning (filmer och serier). För att möjliggöra en textanalys skapade vi en *ground truth* text som är en transkribering av samma talade ljudfil/video framtagen på ett manuellt sätt (se Tabell 1). Denna transkribering utfördes av två forskare som själva kan flytande ryska.

Tabell 1 Antal ord i de två transkriberingarna

	BB-project	*Ground truth*
RBK TV	6 598	6 486
Rossiya-24	5 427	5 940

2.1 Automatiserad transkribering: tal till text

BB-projects webbplats redovisar inte vilka metoder som används för transkribering av tv-kanaler som de erbjuder. Under analysens gång upptäckte vi att transkriberingarna uppvisar hallucinationer och misstag[1] som ofta förknippas med *Whisper Large v3 c Silero VAD* (förkortat till Whisper i följande text). Det är en av de bästa tillgängliga språkmodeller för *tal-till-text* transkribering (Wang m.fl., 2025). Därför antar vi att BB-project är baserat på Whisper modellen. Detta innebär att vår utvärdering av BB-project också ger en inblick i Whisper-modellens effektivitet vid transkribering av det ryska språket.

Whisper, har trots sina styrkor, en del problem som påverkar transkriberingskvaliteten. Dessa problem är kopplade till videofilens format, och tar sig uttryck som hallucinationer beroende på t.ex. träningsdata, bakgrundsljud, icke-mänskliga ljud (t.ex. djur eller vind), mikrofonkvalitet, talarens brytning eller tvekan i röst och andra problem kopplade till t.ex. längden av video/tal (Barański m.fl., 2025; Höhne m.fl., 2025; Macháček m.fl., 2023). För att öka transkriberingstakten kan BB-project behöva klippa videofiler till i mindre delar. Det kan leda till att språkmodellen tappar kontexten och då kan antingen lägga till text som inte uttalats, eller missa delar som faktiskt har uttalats (Barański m.fl., 2025). Andra vanliga problem med Whisper är grammatiska fel och bristande skiljetecken men de är mindre relevanta för just vår analys. Dessa fel kan påverka textens innebörd. Eftersom Whisper-modellen för närvarande är en av de bästa tillgängliga ASR-modellerna för ryska, krävs en balanserad utvärdering för att inte avfärda en potentiellt användbar resurs för studier av politiskt innehåll.

2.2 Metod

En klassisk metod för utvärdering av transkribering är *Word Error Rate* (WER) (Neumann m.fl., 2023). Metoden går ut på att beräkna andelen fel i transkriberingen av en talad ljudfil relativt en *ground truth*. Metoden är att räkna alla substitutioner, borttagningar och insättningar av ord och grammatiska tecken (se formel). Det finns ingen standardnivå på vad som bedöms tillräckligt bra transkribering, men oftast anses under 10% vara bra (Höhne m.fl., 2025).

$$WER = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{N}} \times 100$$

Där:

S = Antal substitutioner

D = Antal borttagningar

I = Antal insättningar

N= Antal ord i *ground truth*

WER är dock en metod som har en del brister när det gäller utvärdering av textens innebördskvalité (Ye-Yi Wang m.fl., 2003). Alla fel vägs lika, även sådana som inte påverkar innebörden. I samhällsvetenskaplig forskning är ofta ett budskap viktigare än den exakta grammatiken (Pentland m.fl., 2023). Därför modifierade vi WER till en mer innehållsfokuserad metod, WERdm *(Word*

[1] Exempel på dessa: 1) algoritmen upprepar siffror och bokstäver, t.ex ”11” blir 1-1-1, och ”ee” blir ”e-e-e”; 2) En återkommande fras som inte sägs i sändningen men finns med i transkriberingen är ”Субтитры создавал DimaTorzok”, (”Undertext skapat av DimaTorzok), i lägen där algoritmen associerar vissa typer av tystnad/ljud med den frasen Det finns viss dokumentation på att detta kommer från träningen av Whisper modellen (*Who is DimaTorzok? #2372*, 2024).

Error Rate Detrimental Meaning)[2]. Den prioriterar fel som påverkar textens innebörd. Nedan beskriver vi alla metodologiska steg för **WERdm**.

1. Manuellt markera ord i *ground truth* texten som påtagligt påverkar textens innebörd
2. Markera alla ord som stavas fel men har viktig domänspecifik betydelse (särskilt namnen på länder, politiska figurer, militära begrepp)
3. Ignorera böjningsfel och fel skiljetecken i den mån dessa inte påverkar innebörden
4. Räkna alla ord som hittats enligt (1) och (2), summera dem
5. Räkna ut skillnaden mellan antalet ord i BB-project transkriberingen och vår egen *ground truth* transkribering.
6. Summera (4) och (5)
7. Dividera 6 med N antal ord i *ground truth* och multiplicera med 100

Denna metod ignorerar därmed många misstag som skulle ha räknats in enligt *WER*, men inte är avgörande för textens innebörd. Vår **WERdm** metod redovisas enligt formeln nedan:

WERdm - word error rate detrimental meaning

$$WERdm = \frac{\mathrm{C} + \mathrm{B} + \mathrm{F}}{\mathrm{N}} \times 100$$

Där

C= skillnaden i antalet ord mellan automatiserad transkribering och *ground truth*

B= Antal felaktiga begrepp, titel, namn

F= fel ord som påverkar innebörden men är inte ett B

N= Antal ord i *ground truth*

[2] Det finns många andra metoder som fokuserar på utvärdering av transkriptions precision, som t.ex. BERTScore. Dessa metoder dock fokuserar främst på träffsäkerhet av alla tecken/ord i transkriberingen, vilket är mindre relevant för samhällsvetenskaplig forskning och vårt specifik syfte (Pentland m.fl., 2023).

3 Resultat

I detta avsnitt presenterar vi problem som identifierades under vår undersökning av BB-project och potentiella lösningar för hantering av dess autogenererade textdata, för vidare analys.

3.1 Problem

Vår analys visar att transkriberingarna från RBK TV och Rossiya-24 har olika egenskaper och kvaliteter. **WERdm** för Rossiya-24 visar 0.09 (9% fel), 0.06 (6% fel) för RBK TV. I Tabell 2 (se nedan) redovisar vi de största faktorerna bakom skillnaderna mellan transkriberingarna, samt problem som karakteriserar både transkriberingen för RBK TV och Rossiya-24 i vårt urval. Observera att vår metod ignorerar grammatiska misstag som böjningsfel och fel skiljetecken vilka förekom i större utsträckning på RBK TV än på Rossiya-24. De fel som bedöms avgörande för FOI:s omvärldsanalys förekom i större utsträckning på Rossiya-24.

Tabell 2 beskrivning av fel.

Problem	**RBK TV**	**Rossiya-24**	**Beskrivning, exempel, i fall av ja.**
Utelämnade stycken/meningar	nej	ja	Några meningar saknades i transkriberingen: t.ex. en mening från Putins tal saknades: ”Varför skulle vi vilja ha en sådan värld om Ryssland inte finns där?[3]”
Hallucinationer	nej	ja	Text som inte uttalats förekommer, t.ex.: ”Undertexter är skapade av DimaTorzok[4]”
Loops (upprepningar)	ja	ja	Upprepning av delar av meningar i transkriberingen men inte i talet
Upprepande teckenserier (siffror/bokstäver)	nej	ja	Upprepning av vissa tecken som inte återkommer i talet t.ex. istället för 11, 1-1-1, istället för ee, e-e-e
Feltranskriberade länder/namn	ja, vissa	ja, vissa	Problemet rör mest Irak och Iran; modellen slumpmässigt ersätter Iran med Irak och vice versa i transkriberingen
Feltranskribering vid blandade språk	ja	ja	Engelska namn/plattformar transkriberas som ryska ord som låter liknande till engelska men har annan betydelse, t.ex. CNN, Truth Social
Felstavat/utelämnat ord inom specialiserade områden (militär)/förkortningar	ja	ja	Militära förkortningar och vissa begrepp, t.ex. rysk förkortning för luftvärnssystem är **ПВО**, som alltid transkriberas fel i vårt urval

Problemens uppkomst kan relateras till flera faktorer som diskuterats i avsnitt 2.1. En central problematisk aspekt är förekomsten av tystnad i ljudfiler. I en studie av Barański m.fl. (2025) resulterade 40 procent av tysta pauser i de analyserade ljudfilerna i AI-hallucinationer. Vidare påverkas kvaliteten på transkriberingarna av ljudkvaliteten på videofiler och val av mikrofoner vid inspelningen liksom av längden på de ljudsegment som används vid klippning. Segment som överstiger 30 sekunder kan exempelvis ge upphov till upprepningar och andra problem (ibid). Felaktiga transkriberingar kan dessutom orsakas av talarens brytning eller talfel, exempelvis läspning.

[3] Författares översättning av en utelämnade del i transkriberingen av Putins tal
[4] Författares översättning av hallucination från Whisper träningsdata: ”Субтитры создавал DimaTorzok” (se diskussion i avsnitt 2.1)

3.2 Möjliga lösningar för hantering av textdata inför analys

De listade felen i Tabell 2 innebär att textanalys av autogenererade transkriberingar från BB-project kräver omfattande förbehandling av data. Nedan redovisas de möjliga verktygen för att förbättra den transkriberade textkvaliteten så de kan bli användbara för vidare analys.

1. *Skapa en bag-of-hallucinations.* För att automatisk hantera de systematiskt förekommande hallucinationer i Whisper-transkriberingar, kan man skapa en så kallad *bag-of-hallucinations* och använda den för textrensning (Barański m.fl., 2025). Det finns redan tillgängliga listor över Whisper hallucinationer för olika språk, vilka kan vidareutvecklas.
2. *Rensa data från loops*: Systematisk borttagning av upprepningar i texten kan göras automatiskt[5] om analysen inte kräver bevarande av verbala upprepningar. Vi antar då att kontexten och innebörden är viktigare för analys än det faktum att vissa ord upprepas.
3. *Manuell granskning för kritiska segment och terminologilistor*: För avsnitt som innehåller viktiga uttalanden, t.ex. tal av politiska ledare rekommenderas manuell kontroll. Vidare ordlistor för militära termer, förkortningar och namn kan användas för detaljerad manuell granskning.

3.3 Återstående problem utan enkel automatiserad lösning

Även om vissa problem kan åtgärdas automatiskt i transkriberade texter, visade det sig att minst tre problem som upptäckts under vår undersökning är mer allvarliga med hänsyn till forskning som bedrivs på FOI, samt är svårare att lösa på ett automatiskt sätt. Dessa är: fel vid blandade språk, särskilt omnämnanden av icke-ryska medieplattformar, t.ex. Truth Social, CNN, osv; felaktig transkribering av förkortningar, t.ex. *förkortningar för luftvärnssystem* (”PVO”, *ПВО* på ryska); och att modellen ofta byter Iran mot Irak och vice versa. Dessa kräver antingen finjustering av modellen med domänspecifika data eller manuell efterbearbetning.

[5] Detta kan göras genom detektering och radering av *loops* med t.ex. *Python*

4 Rekommendationer

Baserat på vår undersökning föreslår vi följande alternativ för användningen av BB-project och liknande autogenererade transkriberingar baserade på Whisper-modellen:

Använda rengjorda transkriberingar för tematisk analys

Vissa ord och begrepp av intresse kan vara feltranskriberade. Analyser som fokuserar på allmänna teman som förekommer i talet är mer lämpliga för BB-project. En metod som kan användas är *topic modelling* (Blei, 2012) som definierar olika teman i stora textmängder genom att räkna ut sannolikheten för att ord förekommer tillsammans i flera dokument. På så sätt kan forskaren identifiera teman som finns i ostrukturerade textdata. Genom att använda den typen av textanalys blir problemet med enskilda feltranskriberade begrepp mindre, andra ord som återkommer samtidigt kan tillräckligt belysa ett tema och ge en analytikern en bra överblick över olika generella teman som återkommer i talen. Observera, att först ska textdata rensas enligt rekommendationer i avsnitt 4. Att fokusera på frekvenser och sannolikheter av ord som förekommer i ett dokument kan inte dock ge en djupare analys av enskilda teman.

Finjustera ASR-modellen med domänspecifika data

Det är tekniskt möjligt att finjustera modellen med användning av egna textdata som innehåller de ord som inte förekommer i de data som Whisper ursprungligen tränades på. Whisper kan bli bättre på att korrekt transkribera tal i en militär kontext om träningsdata innehåller militära termer samt förkortningar som ofta används i media.

Att använda video- och textdata samtidigt

En lovande metod som använder videodata från tv-sändningar beskrivs i en studie av Girbau med flera (Girbau m.fl., 2024). Enligt den metoden kunde forskare med *deep learning*-metoder systematiskt detektera politiska figurers ansikten på japanska och amerikanska tv-kanaler och dra slutsatser om deras ”inre krets” eller nätverk av politiker. Detta är särskilt relevant när vi tar hänsyn till att vissa politiska figurer som deltar i ryska tv-program och debatter ofta tillhör en ”rysk elit”, ”Putins inre krets” (se t.ex. Edel (2023)). Att utveckla den metoden för att övervaka vilka av dessa ”eliter” visas på tv, samt hur ofta och vilka andra som är med skulle vara en bra resurs för säkerhetsanalyser och även för att analysera den politiska utvecklingen i Ryssland. Tillsammans med textanalys kan forskare vidare analysera politiska program, ideologier och narrativ som kan förändras beroende på ”elitens” sammansättning.

5 Slutsats

BB-project och Whisper-transkriberingsmodellen i övrigt för ryska språket kan vara användbara resurser för forskning om Ryssland, men textdata måste behandlas innan de kan analyseras. Vi rekommenderar att textdata kontrolleras med dess videokällmaterial tills autotranskriberingen för det ryska språket förbättras.

6 Referenser

Barański, M., Jasiński, J., Bartolewska, J., Kacprzak, S., Witkowski, M., & Kowalczyk, K. (2025). Investigation of Whisper ASR Hallucinations Induced by Non-Speech Audio. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. https://doi.org/10.1109/ICASSP49660.2025.10890105

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Daukšas, V., Venclauskienė, L., Urbanavičiūtė, K., & Fridman, O. (2024). *War on All Fronts, How the Kremlin's Media Ecosystem Broadcasts the War in Ukraine*. NATO Strategic Communications Centre of Excellence.

Edel, A. (2023, mars 28). *A Day Inside Putin's Surreal Television Empire How the nonstop blare of Russian state media fuels the war effort—And blurs reality.* https://foreignpolicy.com/. https://foreignpolicy.com/2023/05/28/russia-ukraine-war-putin-propaganda-news-media-television/

Girbau, A., Kobayashi, T., Renoust, B., Matsui, Y., & Satoh, S. (2024). Face Detection, Tracking, and Classification from Large-Scale News Archives for Analysis of Key Political Figures. *Political Analysis*, *32*(2), 221–239. https://doi.org/10.1017/pan.2023.33

Höhne, J. K., Lenzner, T., & Claassen, J. (2025). Automatic speech-to-text transcription: Evidence from a smartphone survey with voice answers. *International Journal of Social Research Methodology*, *28*(5), 625–632. https://doi.org/10.1080/13645579.2024.2443633

Macháček, D., Dabre, R., & Bojar, O. (2023). *Turning Whisper into Real-Time Transcription System* (No. arXiv:2307.14743). arXiv. https://doi.org/10.48550/arXiv.2307.14743

Neumann, T. von, Boeddeker, C., Kinoshita, K., Delcroix, M., & Haeb-Umbach, R. (2023). On Word Error Rate Definitions and their Efficient Computation for Multi-Speaker Speech Recognition Systems. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10094784

Pentland, S. J., Fuller, C. M., Spitzley, L. A., & Twitchell, D. P. (2023). Does accuracy matter? Methodological considerations when using automated speech-to-text for social science research. *International Journal of Social Research Methodology*, *26*(6), 661–677. https://doi.org/10.1080/13645579.2022.2087849

Wang, R., Xu, Z., & Lin, F. X. (2025). WhisperFlow: Speech foundation models in real time. *Proceedings of the 23rd Annual International Conference on Mobile Systems, Applications and Services*, 169–182. https://doi.org/10.1145/3711875.3729151

Who is DimaTorzok? #2372. (2024, oktober 4). [Forum]. GitHub/openai/whisper. https://github.com/openai/whisper/discussions/2372

Ye-Yi Wang, Acero, A., & Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, 577–582. https://doi.org/10.1109/ASRU.2003.1318504