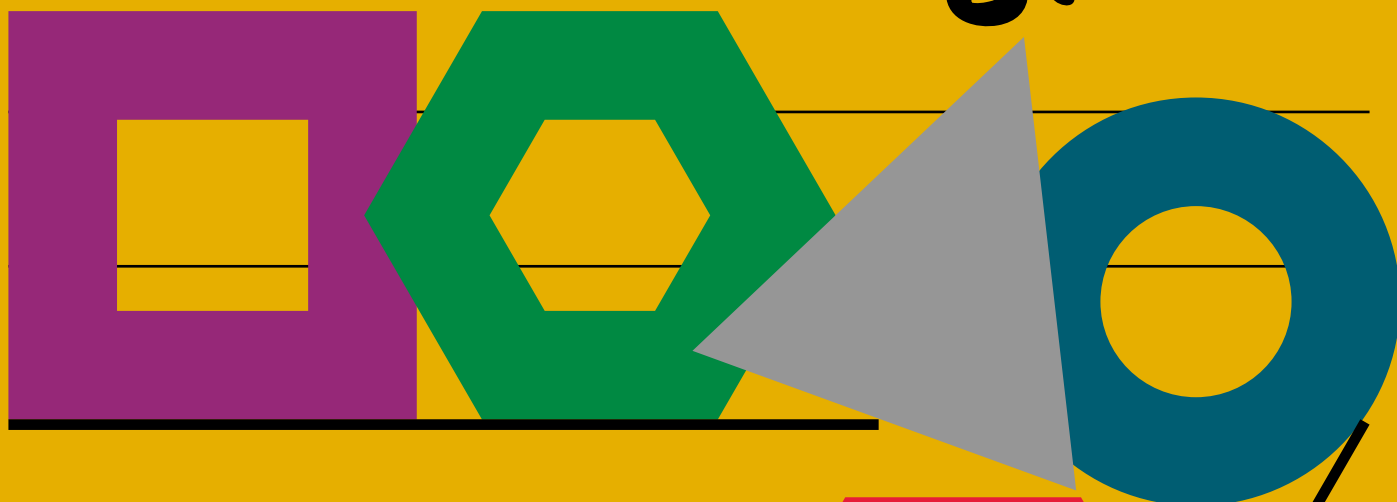




La transcripció automàtica en català, un recurs per a persones amb sordesa? | 2025

Si?



No?



Aquest treball ha estat possible gràcies a:

David Prujà, per la direcció del projecte, i **Silvia Lamela Pérez**, per la recollida del corpus, l'anàlisi de les dades i la redacció de l'informe.

Així mateix, fem una menció especial a totes les persones i entitats que han col·laborat en l'assessorament tecnològic i la facilitació de recursos per a dur a terme el projecte: **Pablo Romero Fresco** i **Luis Alonso Bacigalupe** (GALMA), **Irene De Higes Andino** i **Carla Piqueres Centelles** (TRAMA) i **Jordi Mas** (Softcatalà).

Agraïm també a totes les persones que hi han col·laborat cedint la seva veu per a les mostres: **Ada, Albert, Aline, Ana, Andrea, Anna, Arnau, Diego, Eudal, Facund, Francesc, Guillem, Guim, Guiomar, Javier, Josep, Maïa, Marina, Mariona, Miguel, Núria, Pau, Paula, Robert, Sandra** i **Sara**.

Així mateix, fem un agraïment a totes les persones i entitats que han col·laborat fent difusió del projecte per trobar participants: **Associació d'Amics de la Bressola, El Casal de Perpinyà, Espai Mallorca** i **Helix** (Federació d'Associacions per la Integració de les Persones Sordes Oralistes a la Comunitat Valenciana).

Finalment, agraïm a totes les empreses i entitats desenvolupadores d'aquestes aplicacions que han accedit a posar-se en contacte amb nosaltres i han mostrat interès en el projecte per millorar les seves eines: **Àgils Comunicació, CLiC** i **Google**. Esperem que després de publicar aquest informe en siguin més.

La revisió lingüística ha estat a càrrec de **Jessica Navarro Rosell**.

El disseny i la maquetació del document ha estat a càrrec de **Som Color**.

La impressió ha estat a càrrec de **Gramagraf**.

Índex

01 • Presentació 4

02 • Disseny de la recerca 6

Selecció de les aplicacions	7
Elaboració del corpus	7
Metodologia	8
Anàlisi WER	9
Anàlisi NER	9
Consideracions especials	11

03 • Resultats 12

Viabilitat de la transcripció automàtica com a eina d'accessibilitat i recomanacions d'ús	13
Rendiment de les eines de subtitulació automàtica en directe de les aplicacions de videoconferència (Google Meet i Microsoft Teams)	24
Resultats globals	24
Resultats específics	32
Avaluació de les característiques d'accessibilitat	35
Rendiment de les aplicacions de transcripció en directe i diferit (Speechmatics, AgilText i SCRIBAL)	36
Resultats globals	36
Resultats específics	46
Avaluació de les característiques d'accessibilitat	50
Rendiment de les aplicacions de transcripció de continguts en diferit (YouTube i Softcatalà)	52
Resultats globals	52
Resultats específics	62
Avaluació de les característiques d'accessibilitat	64

04 • Conclusions 66

01 • Presentació

Presentem per tercer any consecutiu una recerca sobre l'estat de la transcripció automàtica en català i la qualitat d'algunes aplicacions concretes. Gràcies al suport del Departament de Política Lingüística, els dos darrers anys l'hem pogut fer de manera més ambiciosa i amb més rigor.

Des de la Federació ACAPPS representem al col·lectiu de persones amb sordesa que es comuniquen oralment, un col·lectiu que sempre ha confiat en la tecnologia per millorar les seves capacitats de comunicació.

L'aparició de fa uns anys d'aplicacions de transcripció automàtica va despertar la curiositat i l'interès de moltes persones amb sordesa, de cara a poder-les fer servir com a eina de suport en la seva comunicació en les situacions que malauradament encara no els són accessibles.

Paral·lelament, algunes empreses amb vocació de millorar l'accessibilitat ens plantejaven si aquestes eines podien resultar útils per fer accessibles la seva activitat i instal·lacions.

És per això que des de la Federació ACAPPS vàrem veure la necessitat de generar un coneixement objectiu sobre la capacitat d'aquestes aplicacions d'esdevenir una eina

d'accessibilitat i, fins i tot, de convertir-se en una mesura de suport a la comunicació homologable al que reclama la legislació d'accessibilitat.

Aquesta és la mirada que impregna tot aquest document; intentar obtenir un criteri objectiu per poder prescriure quines aplicacions poden ser útils en quina situació per a la plena inclusió de les persones amb sordesa.

I amb aquesta mirada hem emprès un repte que va molt més enllà d'aquesta recepta, donat que assumim la responsabilitat de poder generar el coneixement i les metodologies adients per poder avaluar quines aplicacions poden ser més útils en quines situacions —si és que ho arriben a ser— per fer accessible la comunicació oral.

D'aquesta manera, la continuació d'aquest treball implicarà recollir mostres d'àudio de l'activitat ordinària de diferents empreses, institucions i entitats per poder comprovar,

2023

2024

2025

ara sí, amb mostres de comunicació real, si les aplicacions que hem detectat que millor funcionen en català realment funcionen bé en entorns de comunicació autèntics. Aquest és el repte que enfrontarem a partir del 2026.

En aquest sentit, aquesta publicació és la passa prèvia, però que en si mateixa té molt de valor, perquè ja ens permet veure com van millorant les aplicacions, ens ajuda a confirmar algunes de les observacions que fèiem l'any passat sobre les limitacions més acusades que pateixen, i esperem que puguin servir per facilitar la incorporació de mecanismes de millora en els seus processos de desenvolupament.

A les persones amb sordesa, confiem que aquesta publicació les ajudi a poder triar quines aplicacions fer servir en cada cas per facilitar-les la comunicació.

A les empreses, institucions i entitats que volen fer accessible la seva activitat, les convidem a contactar amb nosaltres per ajudar-les a analitzar amb més detall els seus entorns.

A les empreses i grups de recerca que treballen en el desenvolupament d'eines de reconeixement i transcripció de veu, les encoratgem a buscar solucions tècniques a les deficiències que encara tenen aquestes aplicacions quan les fem servir en situacions de comunicació ordinària.

A les persones que estimem la llengua, esperem que aquesta contribució ens ajudi a facilitar-ne l'ús i a garantir els drets lingüístics dels parlants en català; en aquest cas, les persones amb sordesa oralista.

02 • Disseny de la recerca

L'objectiu principal d'aquesta recerca és poder comptar amb dades precises i objectives, recollides seguint una metodologia científica, sobre la qualitat de les transcripcions en català de diverses eines de transcripció automàtica, pensant especialment en la seva efectivitat com a eines d'accessibilitat per a persones amb sordesa que es comuniquen oralment.

Per tant, el disseny de la recerca i l'elecció de les variables ha respost a les necessitats d'accessibilitat d'aquest col·lectiu i als usos que fan de la subtitulació. Així, no s'ha buscat estudiar exclusivament la precisió en el reconeixement dels termes, sinó també el rendiment en situacions de qualitat de so o de captació no ideals —amb soroll de fons o sense microfonar—, les solucions i estratègies que les eines apliquen en casos de barreja de codis o de varietats dialectals allunyades de l'estàndard, o l'ús d'elements propis de la subtitulació per a persones amb sordesa, com són la identificació de parlants.

Així, l'objectiu final d'aquesta recerca és poder reflexionar sobre l'estat actual de les tecnologies de reconeixement de veu i transcripció automàtica en català, i sobre la seva viabilitat real com a eines d'accessibilitat completa per a persones amb sordesa que es comuniquen oralment. Alhora, confiem a poder detectar els punts forts i les mancances de les aplicacions concretes i fer-les saber als seus desenvolupadors, aconseguint així una millora en la seva qualitat o, si més no, un esforç per generar coneixement que pugui desembocar en solucions a les necessitats del nostre col·lectiu. De la mateixa manera, busquem poder dotar els

usuaris d'aquesta tecnologia de criteri tecnològic sobre el funcionament de cadascuna de les aplicacions, perquè puguin escollir-ne les més adients en cada situació i fer-ne un ús conscient.

Selecció de les aplicacions

La selecció de les aplicacions de transcripció automàtica analitzades en aquest estudi s'ha dut a terme a partir de les necessitats i interessos de la base social d'ACAPPS, així com de les entitats que ens demanen assessorament per fer els seus espais i activitats accessibles per a persones amb sordesa que es comuniquen oralment.

En primer lloc, s'han recollit algunes de les consultes més freqüents relatives a l'existència i la qualitat d'eines de transcripció automàtica en català, que tenen a veure sobretot amb l'accessibilitat a converses i activitats en grup, tant en línia com presencials. Així, s'han seleccionat les aplicacions Google Meet i Microsoft Teams, les dues úniques eines de videoconferència que ofereixen subtitulació automàtica en català en el moment d'iniciar la recerca, i AgilText i Speechmatics, dues eines de transcripció en directe en català recomanades per entitats de renom en l'àmbit del reconeixement de veu i l'accessibilitat tecnològica, que han donat bons resultats en el seu ús intern de la Federació ACAPPS.

Alhora, s'ha elaborat i difós una enquesta entre la base social d'ACAPPS, en què s'ha demanat a les persones usuàries quines aplicacions de transcripció automàtica fan servir més i quines voldrien fer servir si en sabessin la qualitat. A partir d'aquesta enquesta, s'ha pogut confirmar que els usuaris fan servir els subtítols de Google Meet i Microsoft Teams per seguir reunions en línia, i s'ha afegit l'eina YouTube, que ha estat la més mencionada en l'accés a subtitulació automàtica de continguts audiovisuals. No obstant això, no hem pogut afegir cap altra aplicació, ja que la majoria de respostes han fet referència a eines que només estan disponibles en castellà.

Finalment, per tal de poder comparar les transcripcions de YouTube amb alguna altra eina de subtitulació en diferit, s'ha seleccionat també el servei de transcripció de Softcatalà, una entitat de referència quant a la promoció de la tecnologia en català amb qui hem col·laborat en ocasions anteriors. Així mateix, s'ha decidit analitzar també, tot i que encara està en proves pilot, l'eina SCRIBAL, que ha estat desenvolupada pel grup de recerca CLiC, amb qui també hem col·laborat anteriorment.

Elaboració del corpus

El corpus utilitzat per a l'elaboració d'aquest estudi està format per 42 mostres, 21 de veus masculines (50 %) i 21 de veus femenines (50 %), que sumen un total de 85 minuts de durada. Totes les mostres han estat recollides específicament per a aquest projecte, amb l'objectiu de poder analitzar cadascuna de les variables contemplades de manera independent. Prèviament a l'enregistrament de les mostres, tots els parlants han signat un document de cessió de les dades de veu.

En primer lloc, s'ha comparat el rendiment de cadascuna de les aplicacions en diverses varietats dialectals. Per fer-ho, s'han enregistrat, en condicions de producció ideals, mostres de contingut idèntic, només amb variacions dialectals de pronúncia, terminacions verbals i lèxic específic. Més concretament, els parlants han llegit en veu alta 19 frases de contingut quotidià, escollides per incloure varietat fonètica i de tipus de sintagmes, en un entorn silenciós i amb una gravadora professional. Aquestes persones han estat parlants de diverses varietats dialectals: central —dues mostres de barceloní i dues de gironí—, balear —dues mostres de mallorquí i dues de menorquí—, nord-occidental —dues mostres de lleidatà i dues d'ebrenc— i valencià —dues mostres de central i dues de meridional—. Malauradament, no s'han pogut trobar participants per incloure una mostra

exhaustiva de varietats dialectals —destaca l'absència del rossellonès, així com de diverses subvarietats de cadascun dels blocs—, però s'ha buscat representativitat pel que fa a quantitat de parlants i a les característiques majoritàries de les variants.

En segon lloc, s'ha comparat el rendiment de cadascuna de les aplicacions amb diversos tipus de discurs. Per fer-ho, s'han enregistrat mostres de les diverses varietats dialectals també en condicions de producció ideals, però de contingut no planejat. Més concretament, els parlants han fet un discurs espontani sobre un tema quotidià —els seus plans per al cap de setmana o l'estiu, què els agrada de la seva ciutat, la darrera pel·lícula que han vist o el darrer llibre que han llegit, en què consisteixen els seus estudis o la seva feina, entre d'altres—, en un entorn silenciós i amb una gravadora professional.

En tercer lloc, s'ha comparat el rendiment de les aplicacions en diversos contextos de producció no ideal, sigui pel context de comunicació o per les característiques del discurs. Per poder aïllar al màxim les variables, aquestes mostres són de parlants de varietat central, la més propera a l'estàndard i la que les aplicacions tendeixen a reconèixer millor.

D'una banda, pel que fa al context de comunicació, s'han gravat mostres amb mala qualitat de so i amb soroll de fons. En el primer cas, s'ha fet servir un enregistrament fet amb un dispositiu mòbil de gamma mitjana, situat a uns metres de distància del parlant, dels mateixos discursos de varietat central dels dos grups de mostres anteriors. D'aquesta manera, el contingut és idèntic excepte pel que fa a la qualitat del so. En el segon cas, s'han enregistrat amb la gravadora professional tant discursos preparats com espontanis de parlants de varietat central, en entorns amb diversos graus i tipus de soroll de fons: trànsit, brunzit de maquinària i altres converses al mateix espai.

D'altra banda, quant a les característiques del discurs, s'han recollit, amb la gravadora

professional i en un entorn silenciós, mostres de converses de dos parlants de varietat central en què es produeixen interrupcions i superposicions de les veus. Finalment, s'han recollit també mostres de parlants de varietat central, en entorn silenciós, amb la gravadora i tant de discurs preparat com espontani, que canvien de llengua entre català i castellà entre frases o dins de la mateixa frase.

Metodologia

Per analitzar les mostres, s'han comparat les transcripcions automàtiques de les eines amb la transcripció manual de cadascun dels textos orals en un doble vessant: quantitatiu, del nombre de diferències, i qualitatiu, de la gravetat dels errors.

La transcripció manual de les mostres ha estat una transcripció literal, és a dir, mantenint tots els errors gramaticals dels parlants, que no s'han considerat errors en la transcripció automàtica. De la mateixa manera, s'han mantingut tots els trets dialectals —lèxics i gramaticals—, en la forma recollida al *Diccionari de la llengua catalana* de l'Institut d'Estudis Catalans (DIEC2) o al *Diccionari català-valencià-balear* (DCVB) d'Alcover i Moll, així com els castellanismes no acceptats, en la seva forma en castellà.

D'altra banda, la puntuació de la transcripció manual ha estat l'essencial i normativa, sense signes de puntuació complexos o estilístics —com el punt i coma o els dos punts fora de les enumeracions—. D'aquesta manera, l'omissió de puntuació estilística, que és més freqüent que l'addició, no s'ha considerat un error en les transcripcions automàtiques.

En el cas de les reformulacions —frases inacabades— s'han marcat amb tres punts suspensius en un únic caràcter, i la nova frase ha començat en majúscula. Les dubitacions —repeticions d'una mateixa paraula— s'han

marcat amb tres punts suspensius en un únic caràcter, i la frase ha seguit en minúscula.

A més, per poder garantir una anàlisi correcta en els processos automàtics, la transcripció manual ha estat en text pla, sense marques tipogràfiques. Aquells sintagmes que en requereixen normativament —usos metalingüístics o títols d'obres, per exemple— han estat marcats amb cometes simples. En les transcripcions automàtiques, també s'han substituït les marques tipogràfiques per cometes simples, de manera que el seu ús no s'ha considerat un error.

De la mateixa manera, el canvi de parlant s'ha indicat amb un guionet i s'han substituït en les transcripcions automàtiques totes les marques de canvi de parlant —guió llarg o altres símbols— per guionet.

Anàlisi WER

El model WER (Word Error Rate), avalua la qualitat de les transcripcions automàtiques pel que fa a la seva desviació respecte del text oral. Més concretament, es calculen el nombre de substitucions, addicions i eliminacions que serien necessàries perquè la transcripció fos idèntica a la manual, i es divideix aquest nombre entre el nombre total de paraules.

En conjunt, un resultat de 0 % implica que les dues transcripcions són idèntiques i no és necessari cap canvi perquè ho siguin. Per tant, com més proper sigui el resultat a zero, més alta serà la qualitat segons aquest model. Orientativament, s'aplica la classificació següent: una precisió inferior al 10 % és excel·lent, entre el 10 % i el 20 % és satisfactòria, entre el 20 % i el 30 % és acceptable i més del 30 % és insuficient.

L'anàlisi WER del corpus s'ha dut a terme a través d'una eina de programació desenvolupada

per Softcatalà, que ens han facilitat gràcies a la col·laboració de totes dues entitats, en dues versions: amb les transcripcions completes i amb una versió normalitzada, sense majúscules ni puntuació. D'aquesta manera, s'ha pogut avaluar tant la desviació real com la desviació del reconeixement de termes, sense tenir en compte aspectes de convencions gràfiques.

Aquesta metodologia permet obtenir una aproximació general de la qualitat de les transcripcions, però els resultats són únicament quantitatius i qualsevol desviació és tractada com un error. Així, aspectes com l'ús de paraules en lloc dels símbols equivalents (*un euro* i *1 €*, per exemple), correccions lingüístiques (*vaig pensar de que* i *vaig pensar que*, per exemple) o diferents estratègies de puntuació ('Porta la llanterna, aquí necessitem més llum' i 'Porta la llanterna. Aquí necessitem més llum', per exemple), fan baixar la puntuació d'una transcripció que no és incorrecta. Per això, aquest estudi complementa els resultats quantitatius WER amb una anàlisi qualitativa NER.

Anàlisi NER

El model NER (Named-Entity Recognition), recollit en la Norma UNE 153010¹, avalua la qualitat de les transcripcions automàtiques pel que fa a la gravetat de la seva desviació respecte del text oral, és a dir, la precisió amb què transmeten les unitats de significat del text oral. Més concretament, s'atorga una penalització a cadascun dels canvis en funció del grau en què afecten el contingut i la coherència:

1 Asociación Española de Normalización y Certificación [AENOR]. (2012).

Subtitulado para personas sordas y personas con discapacidad auditiva (UNE 153010:2012).

Errades menors (-0,25) La persona usuària les percep, però no dificulten la comprensió.
Per exemple:
És més alt que jo.
És mes alt que jo.

Errades estàndard (-0,50) La persona usuària les percep i no li permeten comprendre el significat.
Per exemple:
Demà farà calor.
Damià farà calor.

Errades greus (-1,00) La persona usuària no les percep i li transmeten un significat diferent.
Per exemple:
Eren 100.000 persones.
Eren 5.000 persones.

Reconeixements correctes (sense penalització) Canvis que no constitueixen errors i canvis que es poden considerar encerts.
Per exemple:
Tenim que fer-ho aviat.
Hem de fer-ho aviat.

La suma de totes les penalitzacions es resta del nombre total de paraules, i el resultat es divideix entre el nombre total de paraules per obtenir un percentatge. En conjunt, una precisió del 100 % implica que el contingut del text oral s'ha transmès a la perfecció en la transcripció automàtica. Tanmateix, la puntuació mínima que el model NER considera com a acceptable o suficient per afirmar que el significat original ha arribat al destinatari és del 98 %.

Precisió NER	Equivalència del 0 al 10
< 96 %	0/10
96,4 %	1/10
96,8 %	2/10
97,2 %	3/10
97,6 %	4/10
98,0 %	5/10
98,4 %	6/10
98,8 %	7/10
99,2 %	8/10
99,6 %	9/10
100 %	10/10

L'anàlisi s'ha dut a terme amb l'eina NER Buddy, una aplicació que permet obtenir el percentatge de precisió NER de manera automàtica a través d'un model d'intel·ligència artificial, desenvolupat pel grup de recerca GALMA, de la Universitat de Vigo, i entrenada en català gràcies a la col·laboració amb la nostra entitat. En tots els casos, però, les transcripcions han passat un procés de revisió i validació manual.

Consideracions especials

A més de l'anàlisi amb aquests dos models, s'han avaluat també alguns aspectes de les transcripcions que tenen a veure amb l'accessibilitat més enllà de la qualitat del reconeixement de veu. En aquest sentit, s'ha avaluat si les aplicacions identifiquen els parlants —i de quina manera— i si ofereixen l'opció de personalitzar aspectes tant de visualització —tipografia, color i mida de lletra, posició dels subtítols, etc.— com de característiques de les transcripcions —inclusió de glossaris propis, latència de la transcripció, etc.—.

En aquest sentit, també s'ha considerat que l'accessibilitat implica l'accés —o la possibilitat de tenir-lo a través de paràmetres configurables— a tota la informació del discurs, inclosos aquells elements metalingüístics sobre l'origen i el registre del parlant, de manera que tant l'estandardització de trets dialectals com la traducció dels estrangerismes en la transcripció automàtica ha estat considerada un error lleu.

De la mateixa manera, només s'han considerat reconeixements correctes l'omissió d'elements que no aporten significat, ni lingüístic ni metalingüístic, i que dificulten la lectura o comprensió del discurs escrit —duplicacions de paraules o sintagmes previs a una reformulació, per exemple—. En aquest sentit, la repetició de paraules i la reformulació de frases sense puntuació entre elles —o amb puntuació incorrecta— ha estat considerada un error lleu.

03 • Resultats

A continuació, presentem els resultats de l'anàlisi. Per poder extreure'n conclusions rellevants i fiables, hem avaluat cadascuna de les variables amb les dades que li són més representatives.

La puntuació global es donarà amb dades tant quantitatives (WER) com qualitatives (NER), per tal de tenir-ne una visió global. Aquests resultats conclouen si l'aplicació pot ser una eina d'accessibilitat —és a dir, si pot servir per complir amb les obligacions legals quant a l'accessibilitat a la comunicació de les persones amb sordesa que es comuniquen oralment, en substitució de les eines tradicionals—, si pot ser un recurs de suport a la comunicació —és a dir, una eina d'ús personal per complementar altres mesures d'accessibilitat o per suplir la manca d'accessibilitat en certs contextos— o si no s'hauria de considerar el seu ús per a l'accessibilitat.

Els resultats globals per tipus de discurs i per varietats dialectals, però, es donaran només en puntuació NER, ja que la gravetat dels errors és més representativa de la qualitat de la transcripció. Pel que fa a les situacions de producció no ideals, els resultats es donaran en relació amb els resultats de les mostres equivalents en situacions de producció ideals.

Les puntuacions de les mostres amb mala qualitat de so i soroll de fons seran la diferència en puntuació WER respecte de les mostres de

discurs planejat en varietat central, ja que es tracta de mostres idèntiques en contingut. Per tant, la diferència quantitativa ens pot donar una idea prou representativa de la quantitat d'errors provocats exclusivament per la diferència en el so. En aquest cas, és una puntuació més alta el que suposarà pitjors resultats, però cal recordar que en puntuació WER la valoració canvia cada 10 %.

Les puntuacions de les mostres amb interrupció de parlants i amb barreja de llengües, en canvi, seran la diferència en puntuació NER respecte de les mostres de discurs espontani en varietat central. Per tant, una puntuació lleugerament més baixa suposarà pitjors resultats, ja que en puntuació NER cada 0,4 % equival a 1 punt sobre 10.

En el cas dels resultats específics, es tindran en compte només els resultats de les mostres de discurs preparat i de discurs espontani, i no de les mostres en situacions de producció no ideals, perquè aquestes darreres presenten particularitats que afavoreixen certs tipus d'errors i no permetrien obtenir resultats fidels de la proporció o freqüència de cada tipus d'error en contextos generals.

Així, s'oferirà informació sobre la posició de cada aplicació en un rànquing de nombre d'errors globals i per gravetat. A més, es destacaran els tipus d'error més freqüents de cada eina, així com els que cometin amb una freqüència més alta que la mitjana del conjunt de les eines —si la freqüència del tipus d'error fos igual en totes les aplicacions, seria en cadascuna el 14 % dels errors totals d'aquest tipus—.

Viabilitat de la transcripció automàtica com a eina d'accessibilitat i recomanacions d'ús

Abans de fer una valoració individual de la qualitat de cadascuna de les aplicacions analitzades, considerem adient avaluar el rendiment general d'aquesta tecnologia com a eina d'accessibilitat, no només per poder contextualitzar els resultats individuals, sinó també per fer una primera reflexió sobre la viabilitat d'utilitzar eines

de transcripció automàtica en català com a mesura d'accessibilitat en el nostre context sociolingüístic, i en quines condicions.

Si observem els resultats globals, veiem que la mitjana de les puntuacions del conjunt d'aplicacions queda molt lluny de la puntuació mínima per considerar que les transcripcions automàtiques tenen prou qualitat. Més concretament, la mitjana dels resultats WER, quantitatius, és de 29,18 % (al límit entre acceptable i inacceptable), i els resultats NER, qualitatius, donen una mitjana de 95,24 % (0/10). No obstant això, la diferència entre les eines amb pitjor i millor puntuació és molt gran —de 22,97 % el WER i de 7,85 % el NER—, de manera que els resultats són més aviat indicatius d'una gran heterogeneïtat en la qualitat de les aplicacions de transcripció automàtica analitzades.

En aquest sentit, tot i que les dades no són del tot representatives de l'estat global d'aquesta tecnologia en català —ja que caldria haver fet una anàlisi exhaustiva de tots els productes del mercat—, podem concloure que actualment hi ha diverses empreses que ofereixen un servei d'accessibilitat de molt baixa qualitat, així com un

La mitjana de les puntuacions del conjunt d'aplicacions queda molt lluny de la puntuació mínima per considerar que les transcripcions automàtiques tenen prou qualitat per ser eines d'accessibilitat.

La viabilitat d'utilitzar eines de transcripció automàtica com a mesura d'accessibilitat serà més alta en actes formals, en què els discursos estiguin preparats i assajats, es faci servir un català estàndard i normatiu, hi hagi torns de paraula i el so estigui controlat.

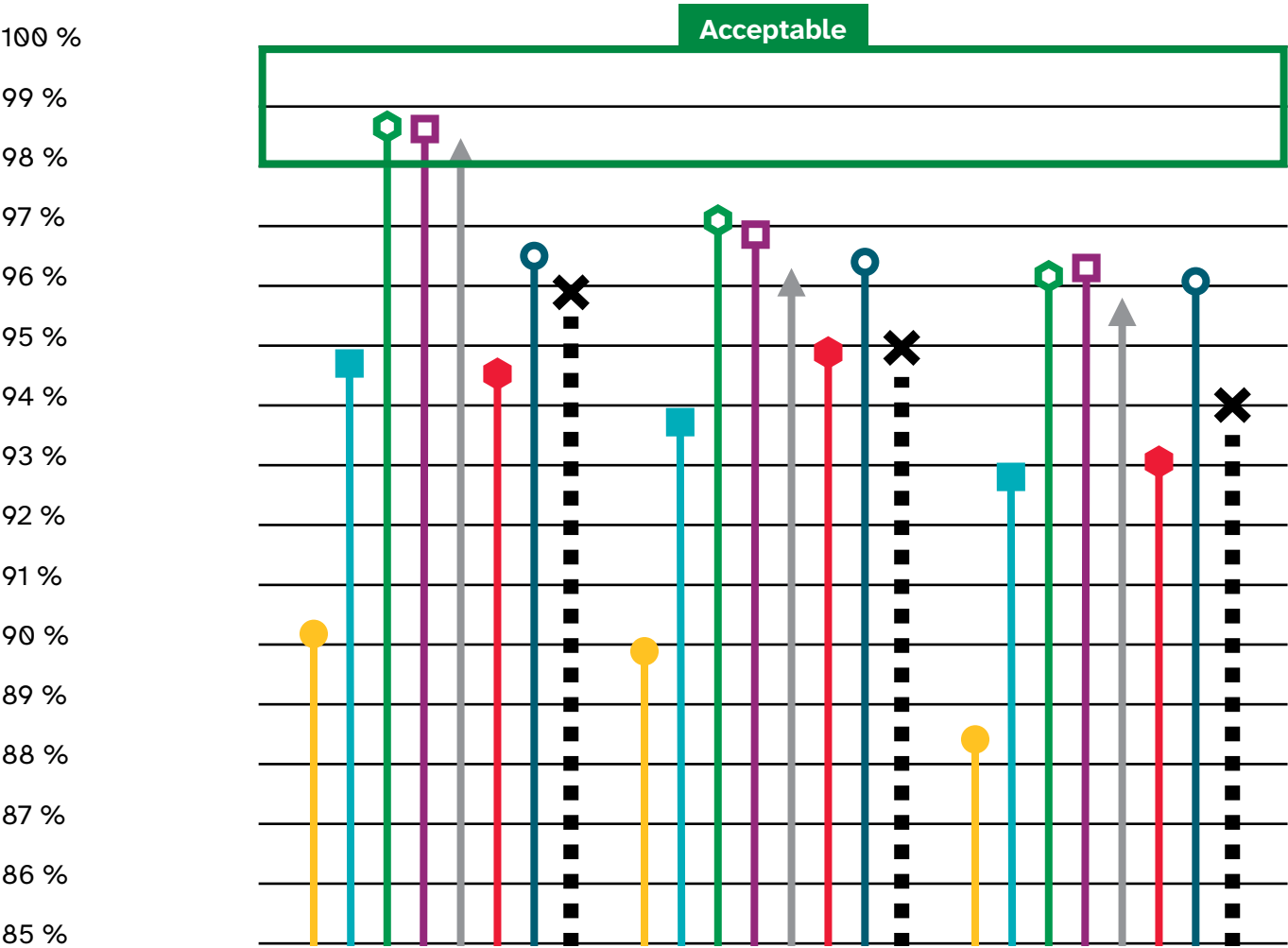
gran nombre de persones amb sordesa i entitats que s'hi comuniquen que les utilitzen de manera habitual, tot i aquesta manca de qualitat.

En una anàlisi més detallada, podem confirmar que les tecnologies de transcripció automàtica tendeixen a donar millors resultats en les mostres de discurs preparat en condicions de producció ideals —amb una mitjana NER de 96,17 %—, seguides del discurs espontani en condicions ideals —95,26 %— i, en darrer lloc, dels discursos en condicions no ideals —94,30 %—, especialment pel que fa a les interrupcions dels parlants i el canvi de llengua entre català i castellà. Per tant, la viabilitat d'utilitzar aquestes eines com a mesura d'accessibilitat real serà més alta en actes formals, en què els discursos estiguin preparats i assajats, es faci servir un català estàndard i normatiu, hi hagi torns de paraula i el so estigui controlat.

Quant a les variants dialectals, confirmem també que el conjunt de les aplicacions donen millors resultats en aquelles més properes a l'estàndard, especialment pel que fa a la pronúncia: central —amb una mitjana NER de

96,22 %—, nord-occidental —96,22 %—, valencià —95,83 %— i, a una distància molt més gran que en la resta de casos, el balear —94,57 %—. Aquest fet cobra sentit en tant que la majoria d'aplicacions de transcripció automàtica no estan pensades com a eines d'accessibilitat, sinó per a productes audiovisuals —en què la llengua acostuma a ser estàndard i l'oralitat és fingida—, o com a complement a altres serveis. Per tant, es confirma que l'ús d'aquestes eines com a mesura d'accessibilitat serà més viable quan la llengua del discurs és estàndard, tant pel que fa a la pronúncia com als aspectes gramaticals i lèxics.

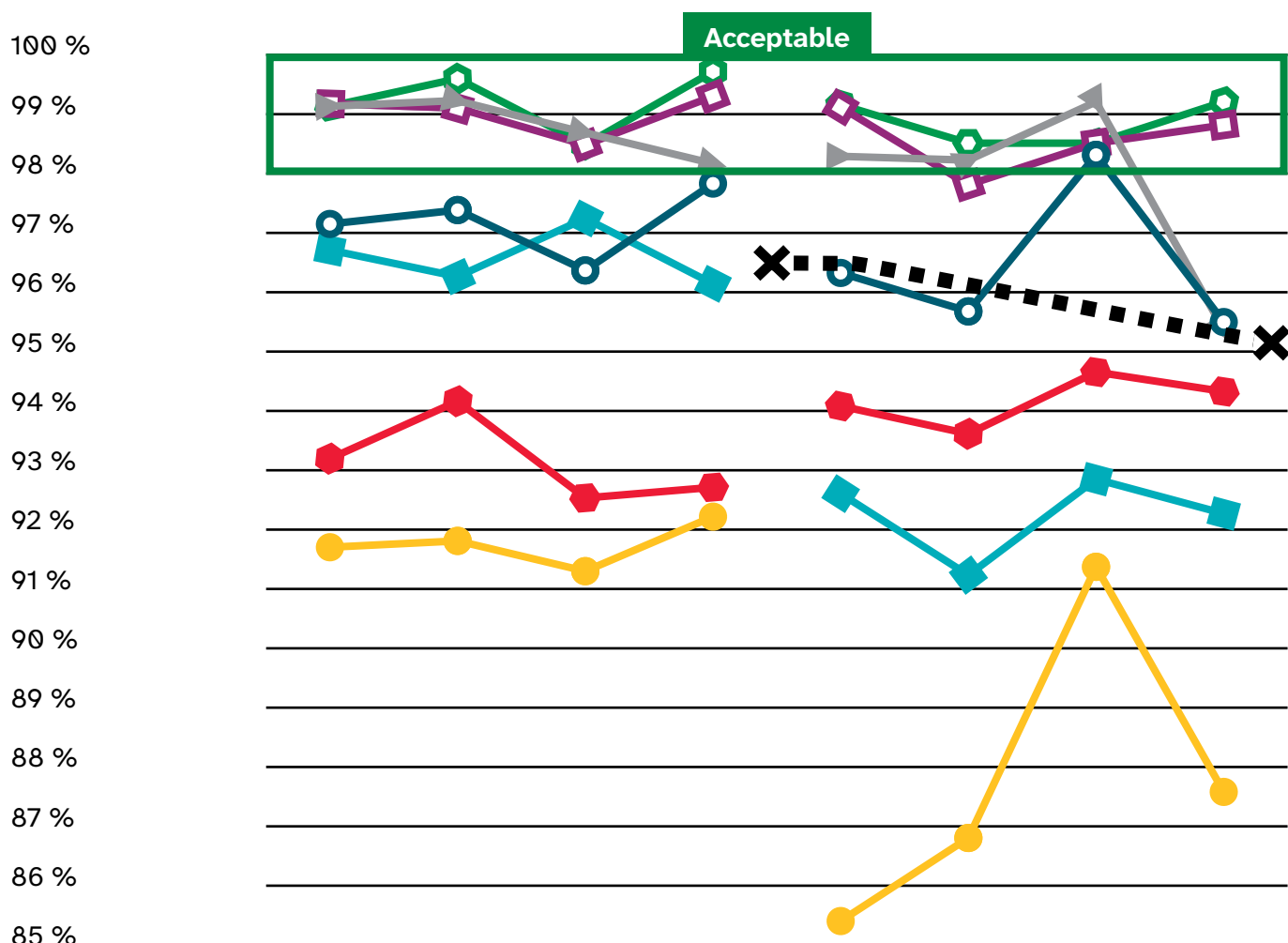
Figura 1. Puntuació NER per tipus de mostra.



	Discurs preparat	Discurs espontani	Discurs o situació no ideal
● Google Meet	90,41 %	90,12 %	88,67 %
■ Microsoft Teams	94,95 %	93,99 %	93,02 %
⬢ Speechmatics	98,94 %	97,38 %	96,43 %
▣ AgilText	98,85 %	97,09 %	96,53 %
▸ SCRIBAL	98,47 %	96,39 %	95,80 %
⬢ YouTube	94,82 %	95,25 %	93,33 %
○ Softcatalà	96,76 %	96,61 %	96,32 %
✕ Mitjana	96,17 %	95,26 %	94,30 %

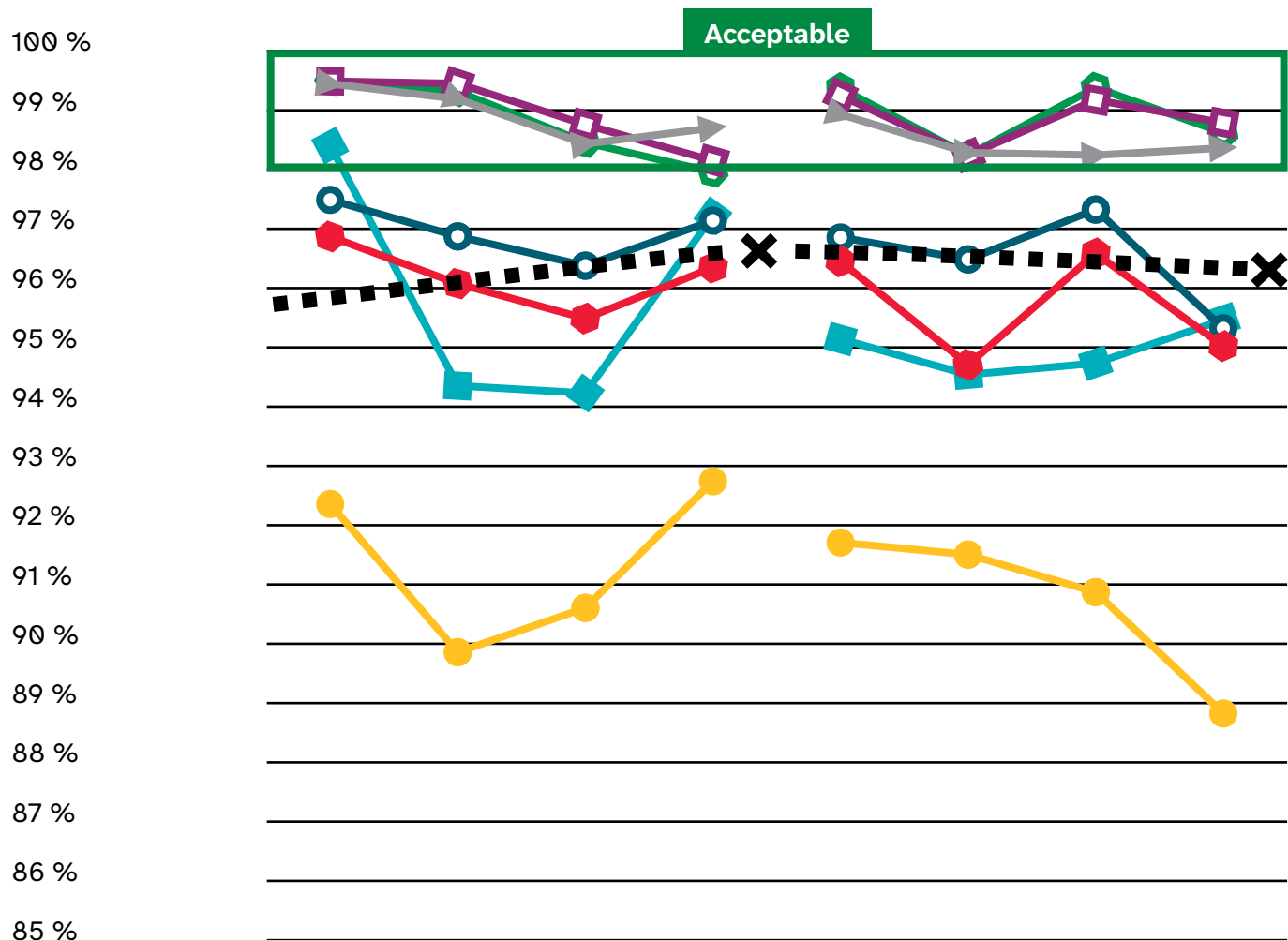
Font: Elaboració pròpia

Figura 2. Puntuació NER dels discursos preparats per varietat dialectal.



	Central				Balear			
● Google Meet	91,71 %	91,86 %	91,37 %	92,21 %	85,40 %	86,77 %	91,29 %	87,50 %
■ Microsoft Teams	96,75 %	96,34 %	97,19 %	96,07 %	92,66 %	91,24 %	92,82 %	92,19 %
⬢ Speechmatics	99,04 %	99,62 %	98,59 %	99,76 %	99,12 %	98,52 %	98,54 %	99,12 %
◻ AgilText	99,20 %	99,12 %	98,49 %	99,31 %	99,13 %	97,87 %	98,53 %	98,87 %
▸ SCRIBAL	99,07 %	99,20 %	98,74 %	98,16 %	98,25 %	98,20 %	99,16 %	95,30 %
● Softcatalà	97,09 %	97,45 %	96,39 %	97,94 %	96,27 %	95,77 %	98,18 %	95,49 %
◆ YouTube	93,13 %	94,08 %	92,59 %	92,68 %	94,04 %	93,72 %	94,72 %	94,34 %
✕ Mitjana del conjunt d'aplicacions	96,54 %				95,11 %			

Font: Elaboració pròpia



	Nord-occidental				València			
● Google Meet	92,34 %	89,79 %	90,51 %	92,76 %	91,76 %	91,53 %	90,90 %	88,92 %
■ Microsoft Teams	98,34 %	94,36 %	94,10 %	97,15 %	95,08 %	94,57 %	94,79 %	95,51 %
◻ Speechmatics	99,49 %	99,28 %	98,41 %	97,98 %	99,36 %	98,18 %	99,36 %	98,67 %
◻ AgilText	99,49 %	99,42 %	98,77 %	98,10 %	99,24 %	98,15 %	99,11 %	98,80 %
► SCRIBAL	99,49 %	99,12 %	98,51 %	98,72 %	98,95 %	98,17 %	98,16 %	98,33 %
● Softcatalà	97,50 %	96,92 %	96,34 %	97,03 %	96,88 %	96,45 %	97,25 %	95,27 %
◻ YouTube	96,94 %	96,02 %	95,48 %	96,37 %	96,47 %	94,83 %	96,68 %	95,00 %
✕ Mitjana del conjunt d'aplicacions	96,74 %				96,30 %			

Pel que fa al gènere, les veus masculines donen uns resultats lleugerament millors en la mitjana de les aplicacions pel que fa als discursos preparats i espontanis, mentre que les veus femenines donen millors resultats en els discursos en situacions no ideals. No obstant això, i tenint en compte que la mostra és reduïda, no podem concloure si es tracta d'una tendència o si és fruit de la casualitat.

Si revisem els resultats de la recerca rellevant en el camp, trobem una gran heterogeneïtat; diversos estudis destaquen millors resultats en veus masculines, mentre que d'altres conclouen millors resultats en veus femenines o, fins i tot, l'absència de biaix de gènere². En general, però, els estudis coincideixen en el fet que la presència de biaix de gènere dependrà de les condicions de disseny, creació i entrenament de cadascun dels models concrets —en aspectes com la selecció de les mostres, l'anotació de les dades o les representacions intermèdies—³.

Per tant, en aquest informe no podem oferir resultats globals per a aquesta variable, i caldria avaluar-la de manera més extensa en cadascuna de les aplicacions analitzades. No obstant això, la poca diferència entre els resultats de veus masculines i femenines dels tres grups de mostres, i el fet que no doni millors resultats sempre el mateix gènere, és indicatiu que globalment existeix la possibilitat de fer servir aquesta tecnologia en català sense biaix.

Si tenim en compte la gravetat dels errors, el 63 % del conjunt d'errors són lleus, és a dir, detectables per a l'usuari i que no afecten el contingut. En general, es tracta d'errors de puntuació —punts i a part entre paraules d'un mateix sintagma o comes entre subjecte i

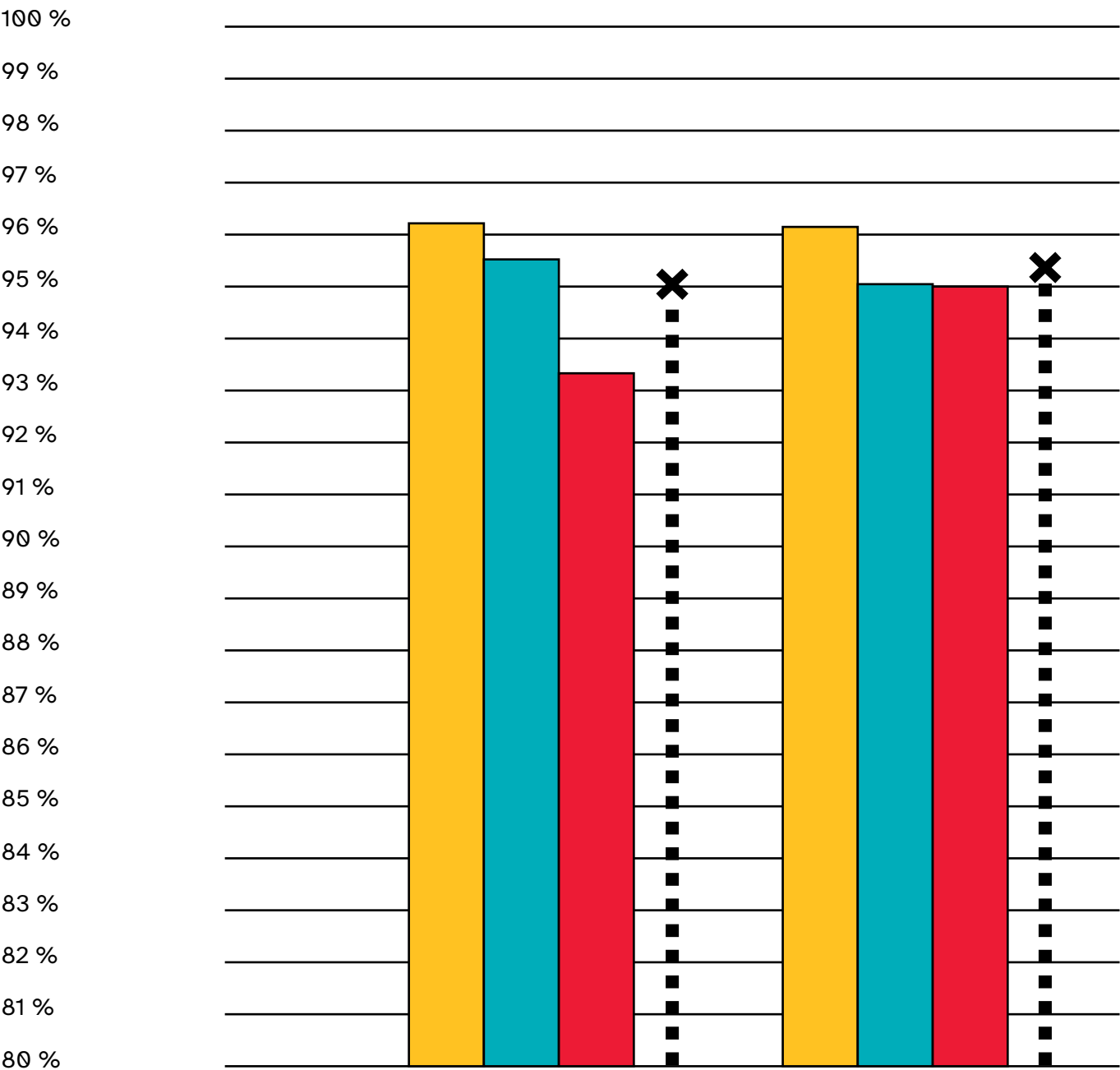
predicat, entre d'altres— o de petites omissions —d'articles definits o de preposicions regides pel verb, entre d'altres—, tot i que també trobem errors de reconeixement —manca de concordança en nombre o gènere entre determinants, noms, adjectius i verbs, entre d'altres—.

Un 32 % dels errors són estàndard, és a dir, detectables per a l'usuari i que impedeixen accedir al contingut. En gairebé tots els casos es tracta d'errors de reconeixement, ja sigui l'ús d'una paraula diferent de l'original o d'un conjunt de caràcters sense significat en català.

Finalment, un 5 % dels errors són greus, és a dir, indetectables per a l'usuari i que fan que accedeixi a informació errònia. Tot i que és el percentatge més baix, són el tipus d'error més perillós, perquè poden donar una falsa sensació de qualitat i accessibilitat, quan en realitat no s'està accedint a la informació real. Aquesta mena d'errors acostumen a ser de puntuació —omissió de puntuació abans o després de partícules negatives, invertint el sentit de l'oració, o canvi de signes d'interrogació per punts i viceversa, entre d'altres—, tot i que també en trobem de reconeixement, especialment pel que fa a xifres, i, en alguns casos, omissions de frases o fragments de text.

-
- 2 Feng, S., Halpern, B. M., Kudina, O. i Scharenborg, O. (2024). Towards inclusive automatic speech recognition, *Computer Speech & Language*, 84. i Garnerin, M., Rossato, S. i Besacier, L. (2021). *Investigating the Impact of Gender Representation in ASR Training Data: a Case Study on Librispeech*. 3rd Workshop on Gender Bias in Natural Language Processing.
- 3 Garnerin, M., Rossato, S. i Besacier, L. (2021). *Investigating the Impact of Gender Representation in ASR Training Data: a Case Study on Librispeech*. 3rd Workshop on Gender Bias in Natural Language Processing.

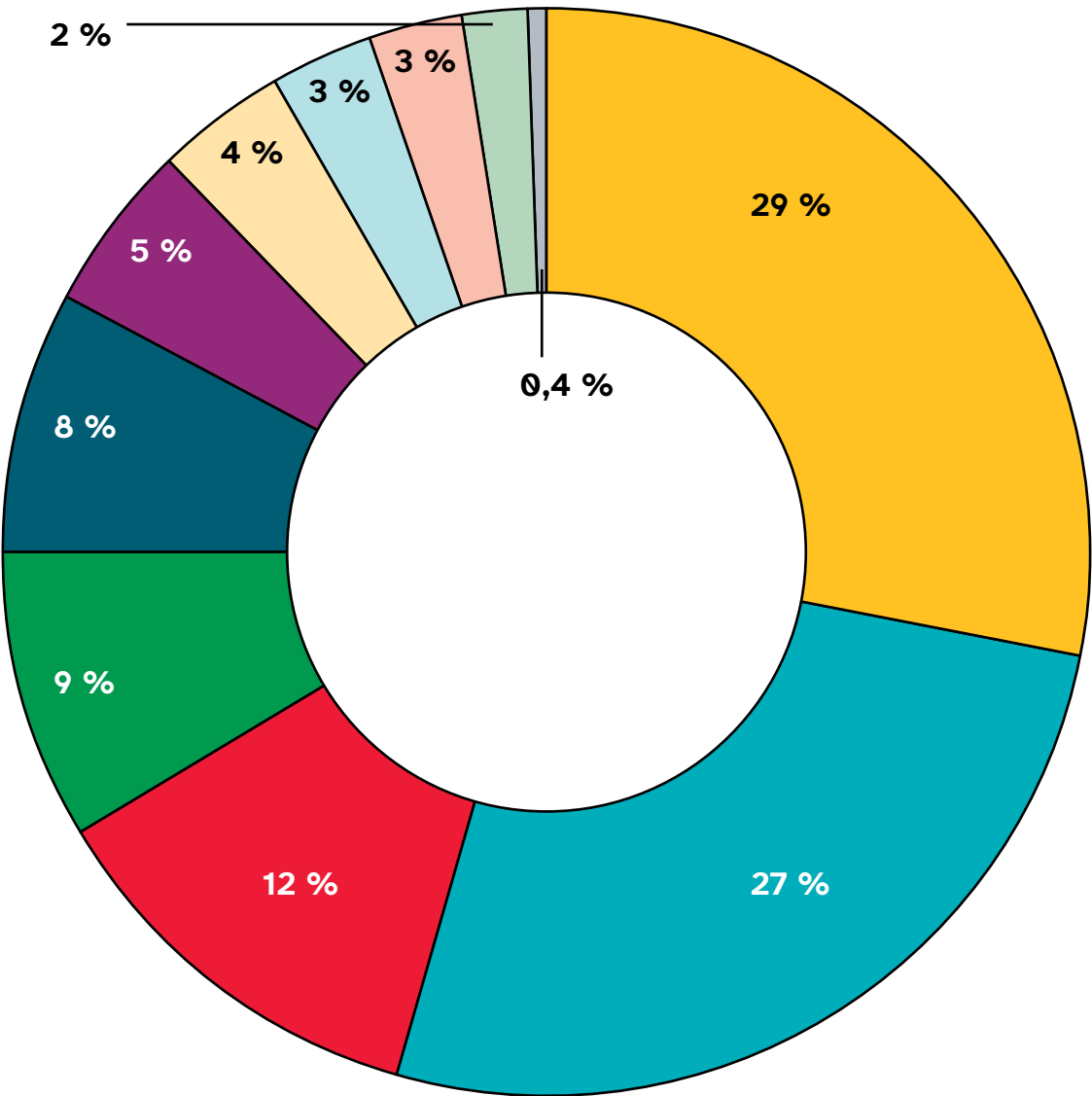
Figura 3. Puntuació NER per gènere.



	Masculí	Femení
Discurs preparat	96,22 %	96,12 %
Discurs espontani	95,44 %	95,07 %
Discurs o situació no ideal	93,43 %	95,00 %
Mitjana	95,03 %	95,39 %

Font: Elaboració pròpia

Figura 4. Tipus d'errors més freqüents en els discursos preparats i espontanis del conjunt de les aplicacions.



Reconeixement sense causa aparent	Estandardització dialectal
Puntuació	Noms propis
Convencions ortogràfiques	Addicions i al·lucinacions
Omissions	Castellanismes
Dubitacions i reformulacions	Xifres
Canvis de temps verbals	

Font: Elaboració pròpia

Els tipus d'error més freqüents en el conjunt de les mostres han estat els de reconeixement erroni de termes sense causa aparent i els de puntuació.

Quant al tipus d'errors, els més freqüents en el conjunt de les mostres —29 % dels errors totals— han estat els de reconeixement erroni de termes sense causa aparent, és a dir, l'ús d'una paraula diferent de l'original —o, en alguns casos, d'un conjunt de caràcters sense significat— en un context del discurs que no presenta cap particularitat.

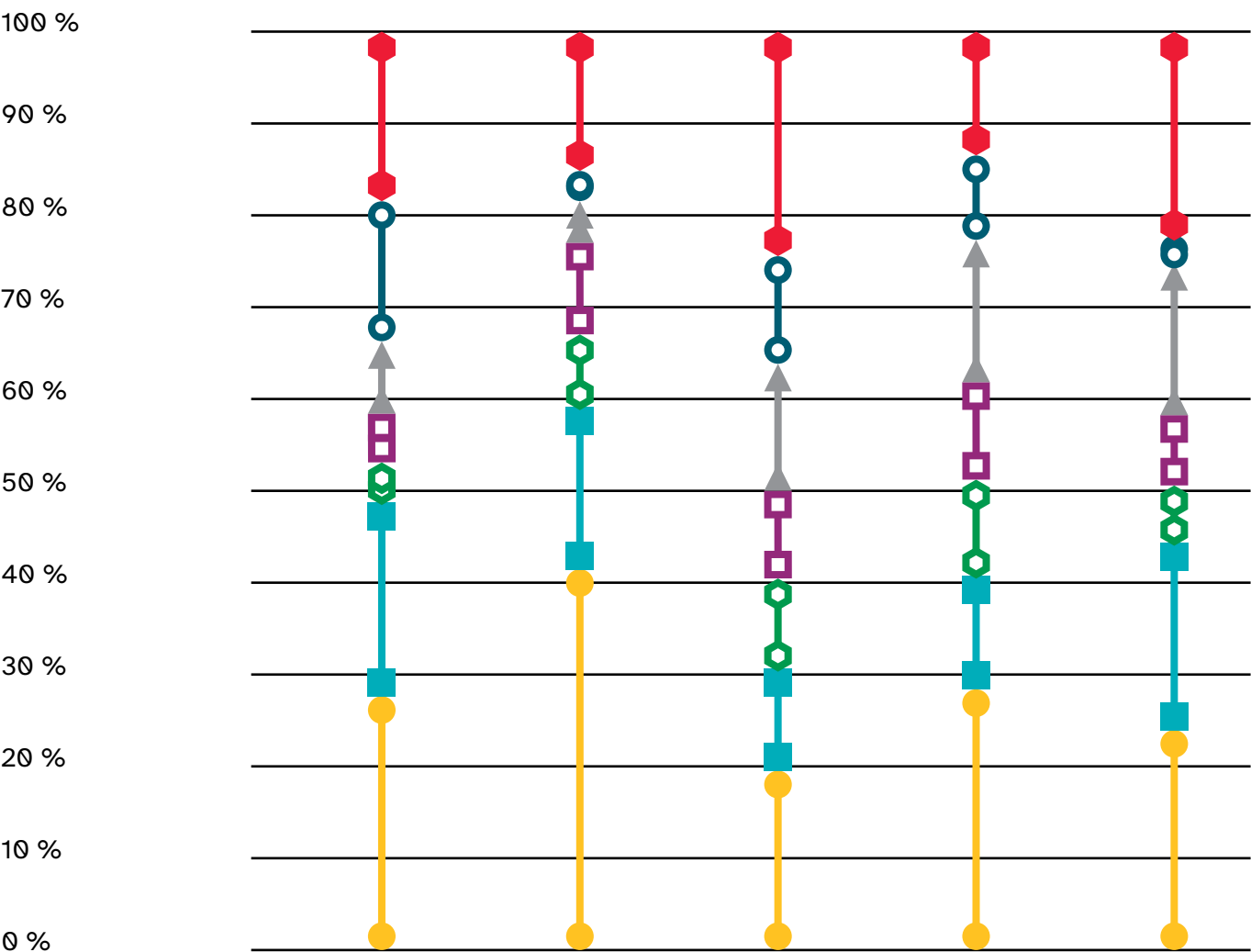
El segon tipus d'errors més freqüents —27 % dels errors totals— tenen a veure amb la puntuació. Com s'ha mencionat anteriorment, en molts casos es tracta d'errors lleus relacionats amb l'ús de puntuació entre paraules d'un mateix sintagma o oració i d'incompliments de la normativa i convencions —coma entre subjecte i predicat o absència de cometes per a cites directes, per exemple—.

En aquest sentit, també trobem altres errors pel que fa a convencions —12 % dels errors totals—, en l'ús de majúscules —majúscules després de comes o minúscules després de punts, però també minúscules per a noms propis o majúscules per a noms comuns— o en faltes d'ortografia relacionades amb aspectes com l'apostrofació i l'ús d'accents diacrítics.

Aquests tres tipus d'errors, en general, depenen exclusivament de l'entrenament de les aplicacions, i no de les característiques del discurs, tot i que tendeixen a ser menys freqüents quan el parlant té

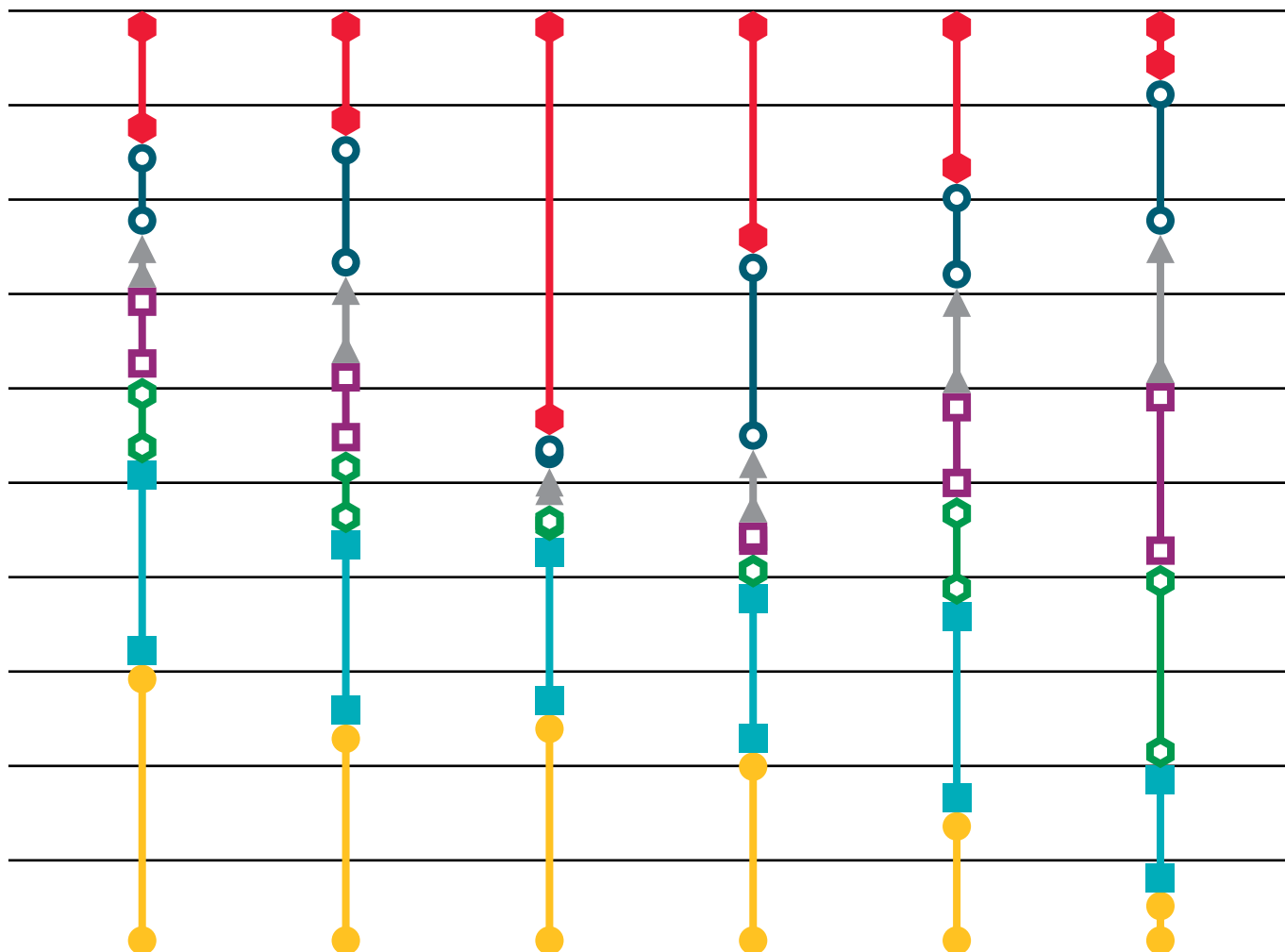
un ritme més pausat o una dicció més clara. Això no obstant, sí que s'han pogut detectar alguns aspectes del discurs en què tendeixen a aparèixer més problemes de reconeixement que en d'altres: dubitacions o reformulacions —paraules repetides, frases inacabades i mots crossa—, noms propis —especialment antropònims i noms d'entitats, però també topònims i noms de festivitats—, paraules d'altres llengües —sobretot castellanismes no acceptats— i xifres —especialment quan es tracta d'un seguit de números—.

Figura 5. Percentatge dels errors de cada tipus de cadascuna de les aplicacions.



	Reconeixement sense causa aparent	Puntuació	Addicions i al·lucinacions	Omissions	Dubitacions i reformulacions
YouTube	18 %	14 %	24 %	13 %	23 %
Softcatalà	15 %	4 %	12 %	9 %	1 %
SCRIBAL	8 %	5 %	14 %	16 %	17 %
AgilText	6 %	10 %	9 %	11 %	8 %
Speechmatics	5 %	8 %	10 %	10 %	6 %
Microsoft Teams	21 %	17 %	11 %	12 %	20 %
Google Meet	28 %	42 %	20 %	29 %	25 %

Font: Elaboració pròpia



Convencions ortogràfiques	Canvis de temps verbals	Xifres	Noms propis	Castellanismes	Estandardització dialectal
14 %	13 %	44 %	25 %	18 %	7 %
10 %	15 %	4 %	21 %	11 %	16 %
6 %	9 %	4 %	7 %	11 %	16 %
9 %	9 %	0 %	3 %	11 %	19 %
9 %	9 %	4 %	3 %	11 %	21 %
21 %	21 %	19 %	18 %	22 %	13 %
31 %	25 %	26 %	22 %	16 %	7 %

Rendiment de les eines de subtitulació automàtica en directe de les aplicacions de videoconferència (Google Meet i Microsoft Teams)

Resultats globals

En termes generals, la puntuació global de totes dues aplicacions queda molt lluny de la puntuació mínima per considerar que presenten prou qualitat per ser eines d'accessibilitat per a persones amb sordesa en compliment de la legislació, i tampoc no podem recomanar fer-ne ús com a eines de suport en casos de manca d'accessibilitat.

Concretament, Google Meet presenta la pitjor puntuació de les set aplicacions analitzades, amb una mitjana WER de 45,65 % (insatisfactòria) i una mitjana NER de 89,73 % (0/10). Microsoft Teams, per la seva banda, presenta la segona pitjor puntuació de totes les aplicacions, amb una mitjana WER de 32,43 % (insuficient) i una mitjana NER de 93,99 % (0/10).

Com a primera conclusió, per tant, entenem que actualment no hi ha cap aplicació de

videoconferència que ofereixi subtitulació automàtica en català amb una qualitat suficient per a considerar que garanteix l'accessibilitat de persones amb sordesa.

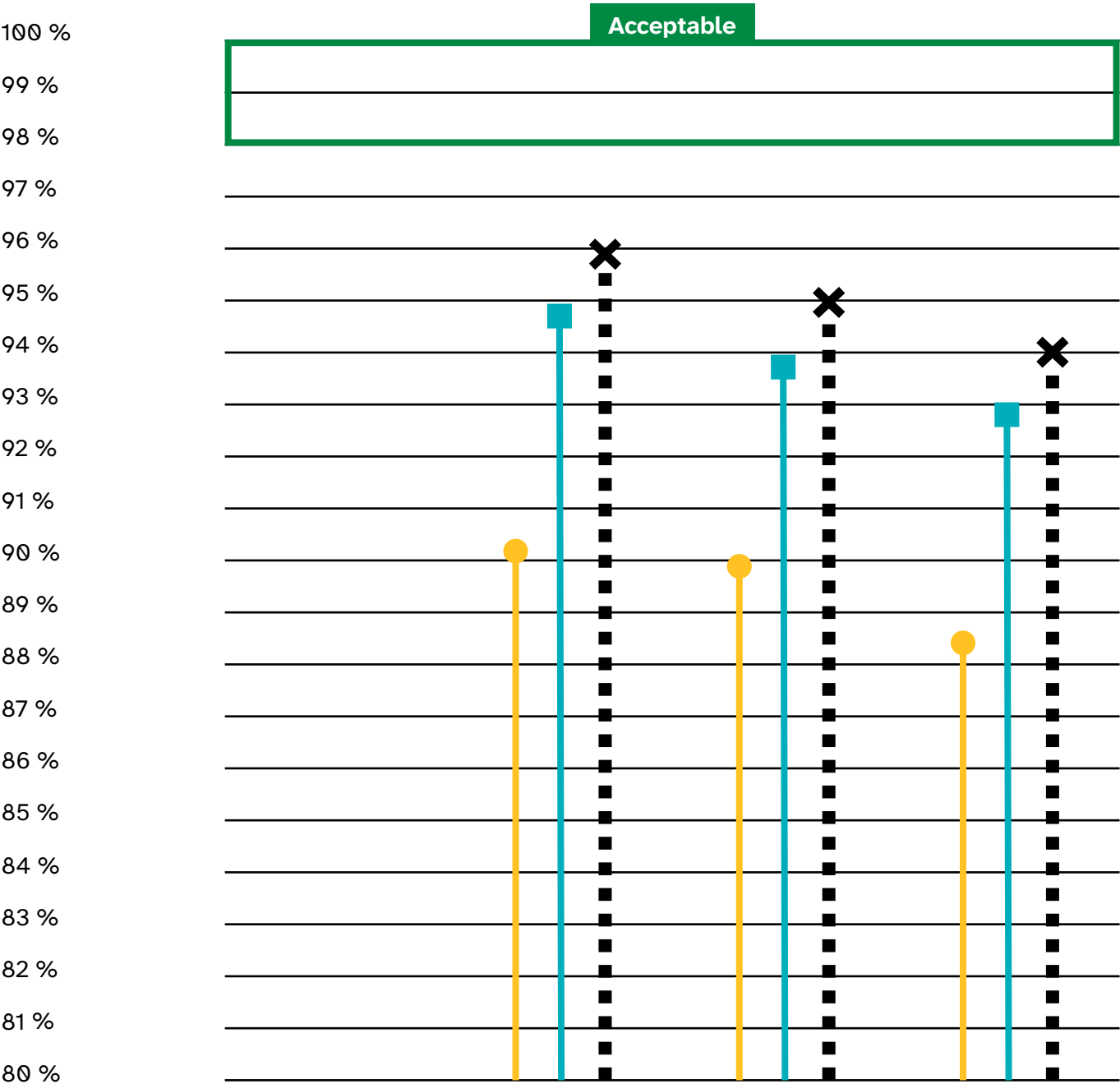
Per tipus de mostra, Google Meet presenta resultats similars en els discursos preparats —NER 90,41 %— i espontanis —NER 90,12 %—, però molt més baixos en condicions de producció no ideals —NER 88,67 %—.

Més concretament, presenta grans diferències de puntuació entre varietats dialectals. Els millors resultats són en dialecte central —NER de discurs preparat de 91,79 %—, seguits pel nord-occidental —91,35 %—, el valencià —90,78 %— i, molt allunyat, el balear —87,74 %—. Les puntuacions del discurs espontani són semblants.

Quant a les situacions de comunicació no ideals, els pitjors resultats són en les mostres amb interrupcions dels parlants —NER més baixa en 7,94 % que els discursos espontanis del mateix dialecte en condicions ideals de producció—, seguits de les de canvi de llengua —NER més baixa en 2,17 %—, les de soroll de fons —WER més alta en 6,05 % que el mateix discurs en situació ideal— i les de mala qualitat —WER més alta en 4,92 %—.

Google Meet presenta la pitjor puntuació de les set aplicacions analitzades, i Microsoft Teams presenta la segona pitjor puntuació de totes les aplicacions.

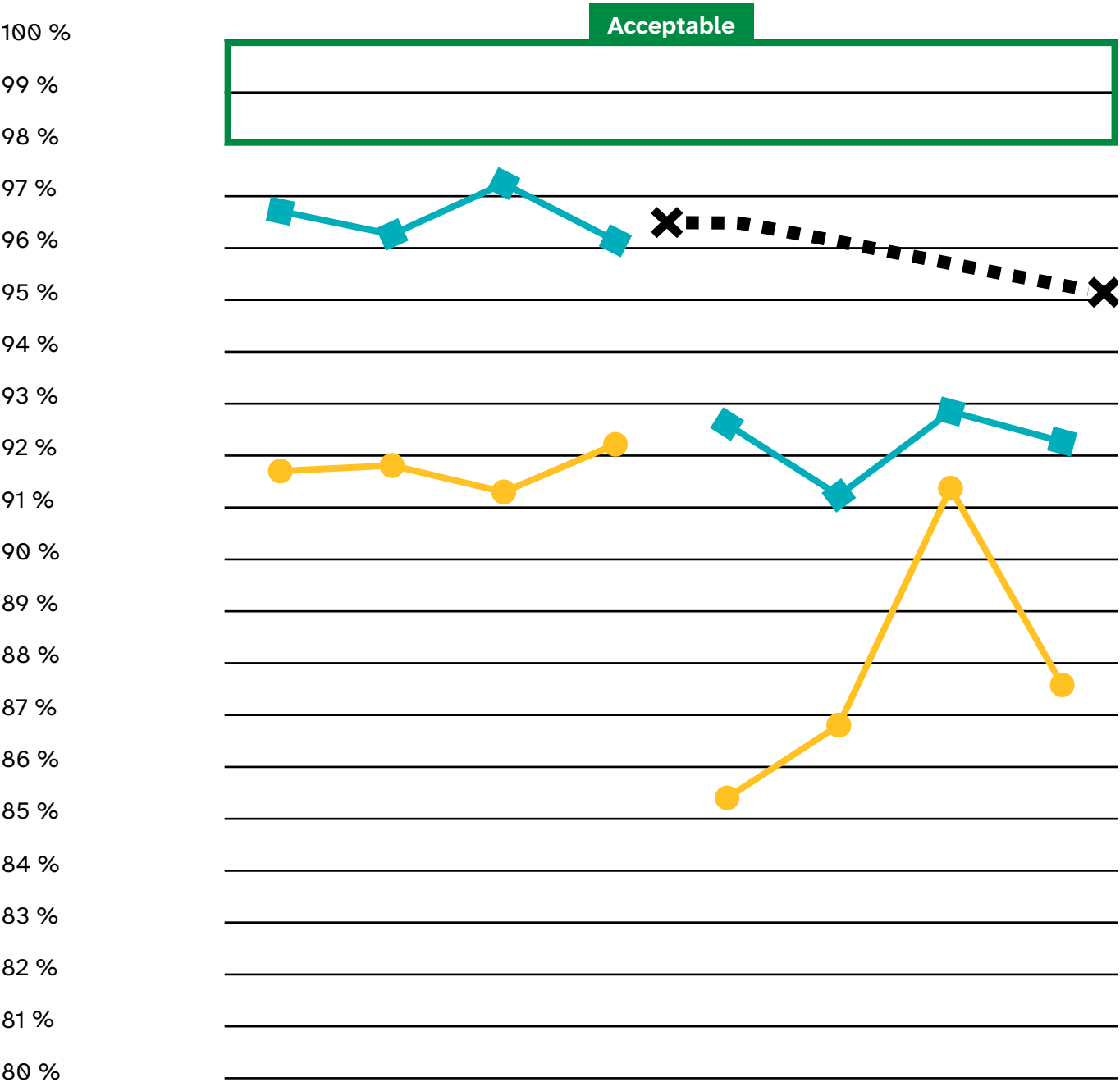
Figura 6. Puntuació NER per tipus de mostra.



	Discurs preparat	Discurs espontani	Discurs o situació no ideal
● Google Meet	90,41 %	90,12 %	88,67 %
■ Microsoft Teams	94,95 %	93,99 %	93,02 %
✕ Mitjana del conjunt d'aplicacions	96,17 %	95,26 %	94,30 %

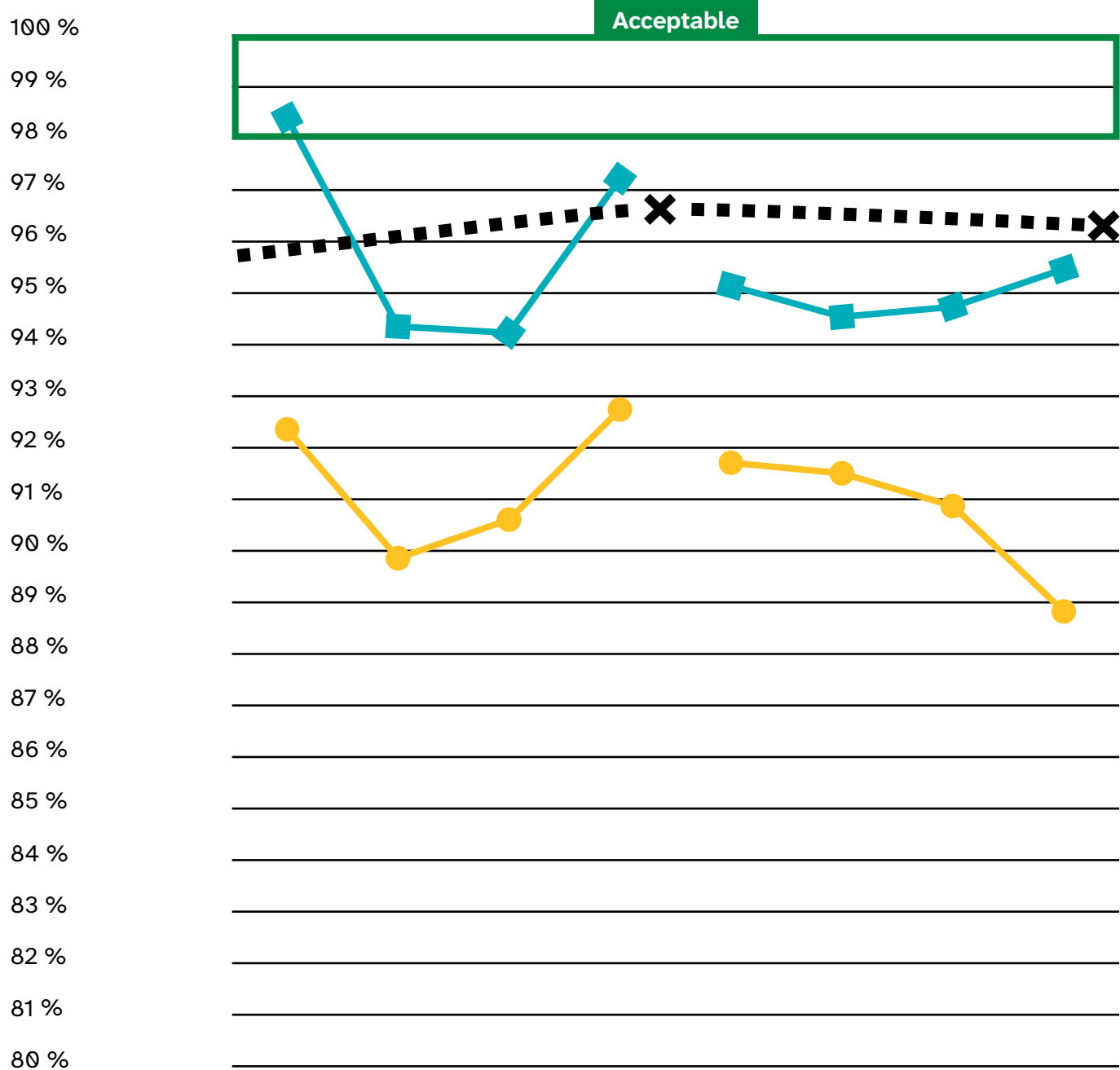
Font: Elaboració pròpia

Figura 7. Puntuació NER dels discursos preparats per varietat dialectal.



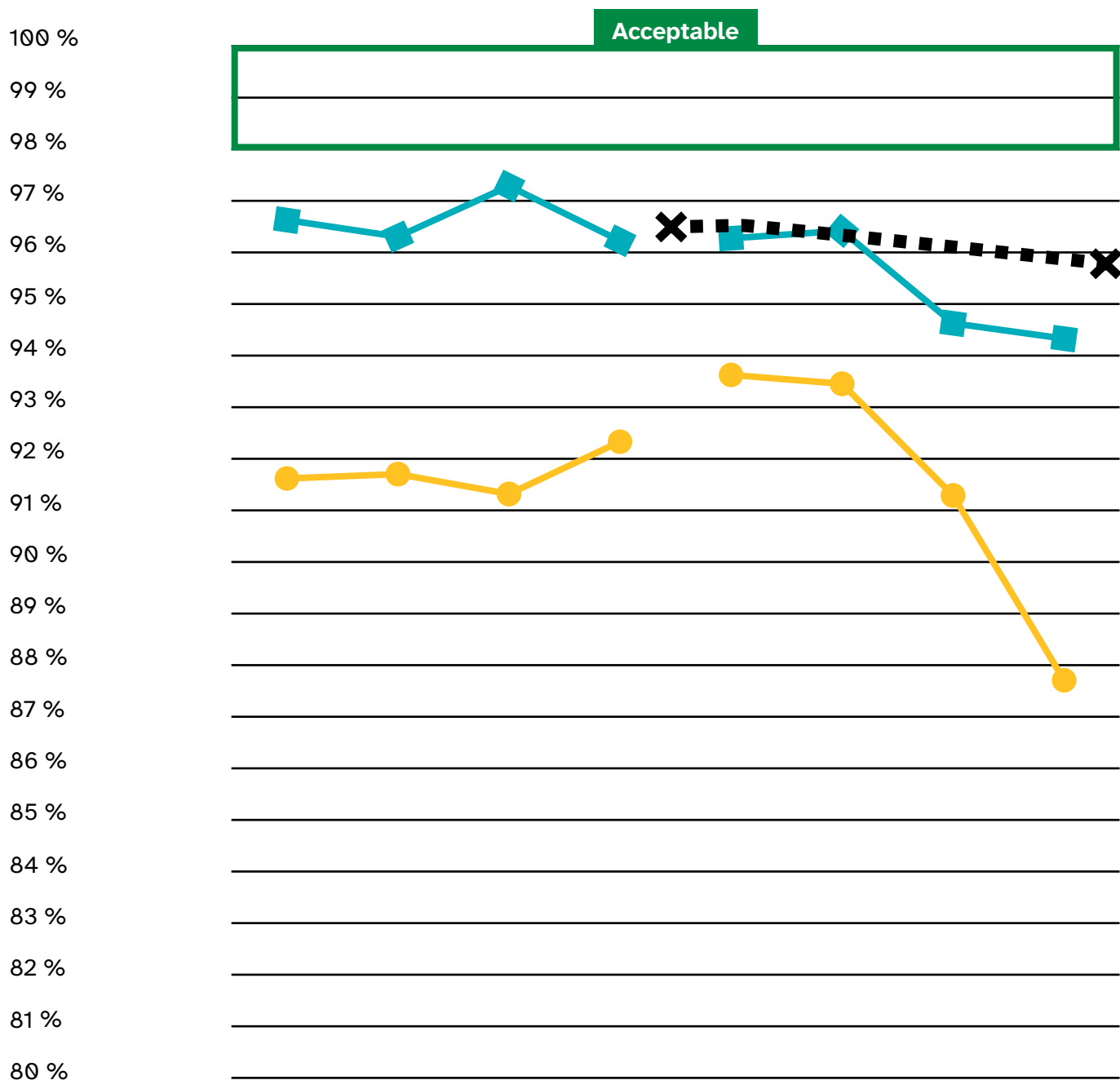
	Central				Balear			
● Google Meet	91,71 %	91,86 %	91,37 %	92,21 %	85,40 %	86,77 %	91,29 %	87,50 %
■ Microsoft Teams	96,75 %	96,34 %	97,19 %	96,07 %	92,66 %	91,24 %	92,82 %	92,19 %
✕ Mitjana del conjunt d'aplicacions	96,54 %				95,11 %			

Font: Elaboració pròpia



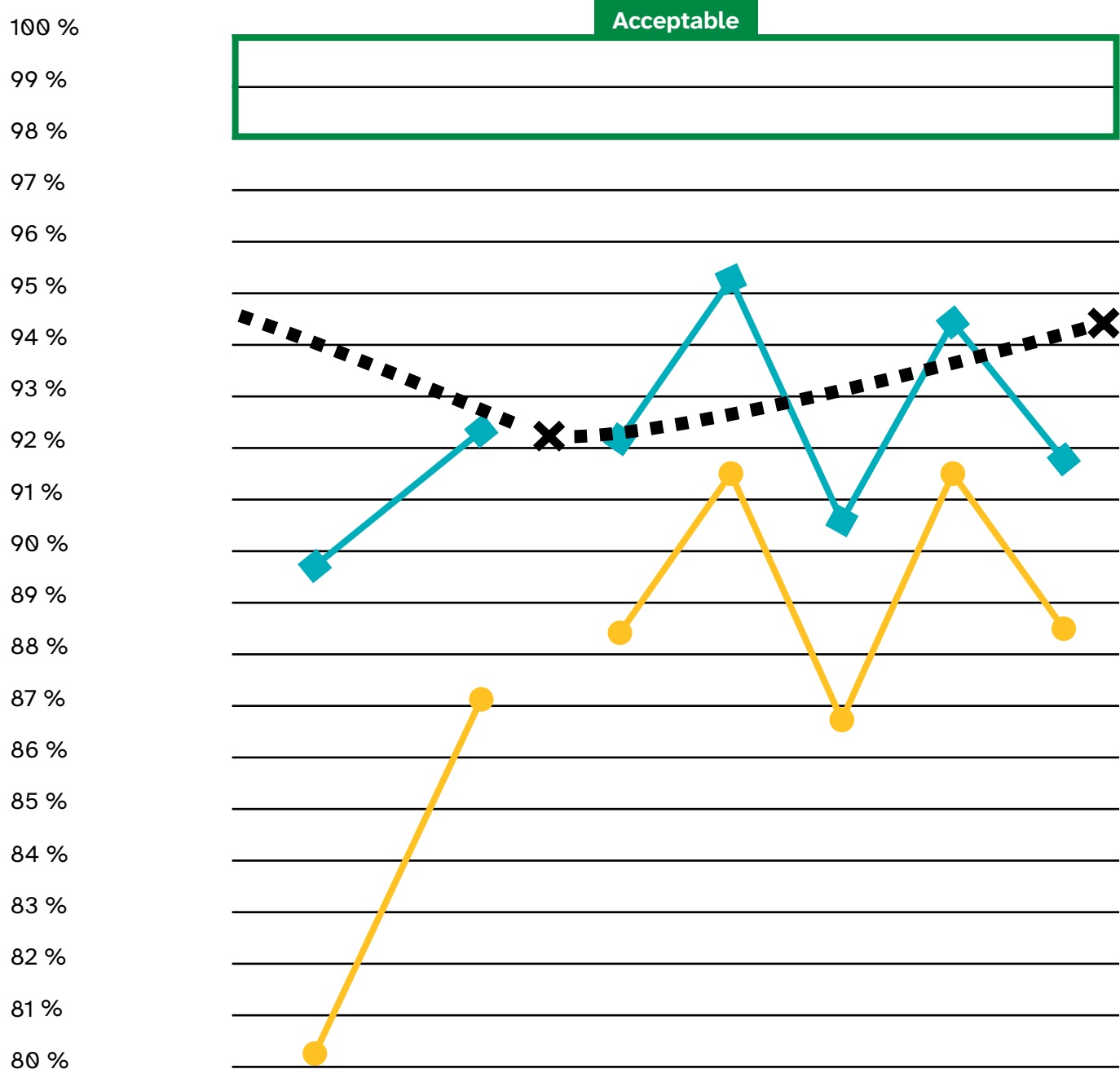
	Nord-occidental				València			
● Google Meet	92,34 %	89,79 %	90,51 %	92,76 %	91,76 %	91,53 %	90,90 %	88,92 %
■ Microsoft Teams	98,34 %	94,36 %	94,10 %	97,15 %	95,08 %	94,57 %	94,79 %	95,51 %
✕ Mitjana del conjunt d'aplicacions	96,74 %				96,30 %			

Figura 8. Puntuació NER per tipus de discurs.



	Central - preparat				Central - espontani			
● Google Meet	91,71 %	91,86 %	91,37 %	92,21 %	93,73 %	93,41 %	91,24 %	87,80 %
■ Microsoft Teams	96,75 %	96,34 %	97,19 %	96,07 %	96,17 %	96,41 %	94,78 %	94,44 %
✕ Mitjana del conjunt d'aplicacions	96,54 %				95,91 %			

Font: Elaboració pròpia



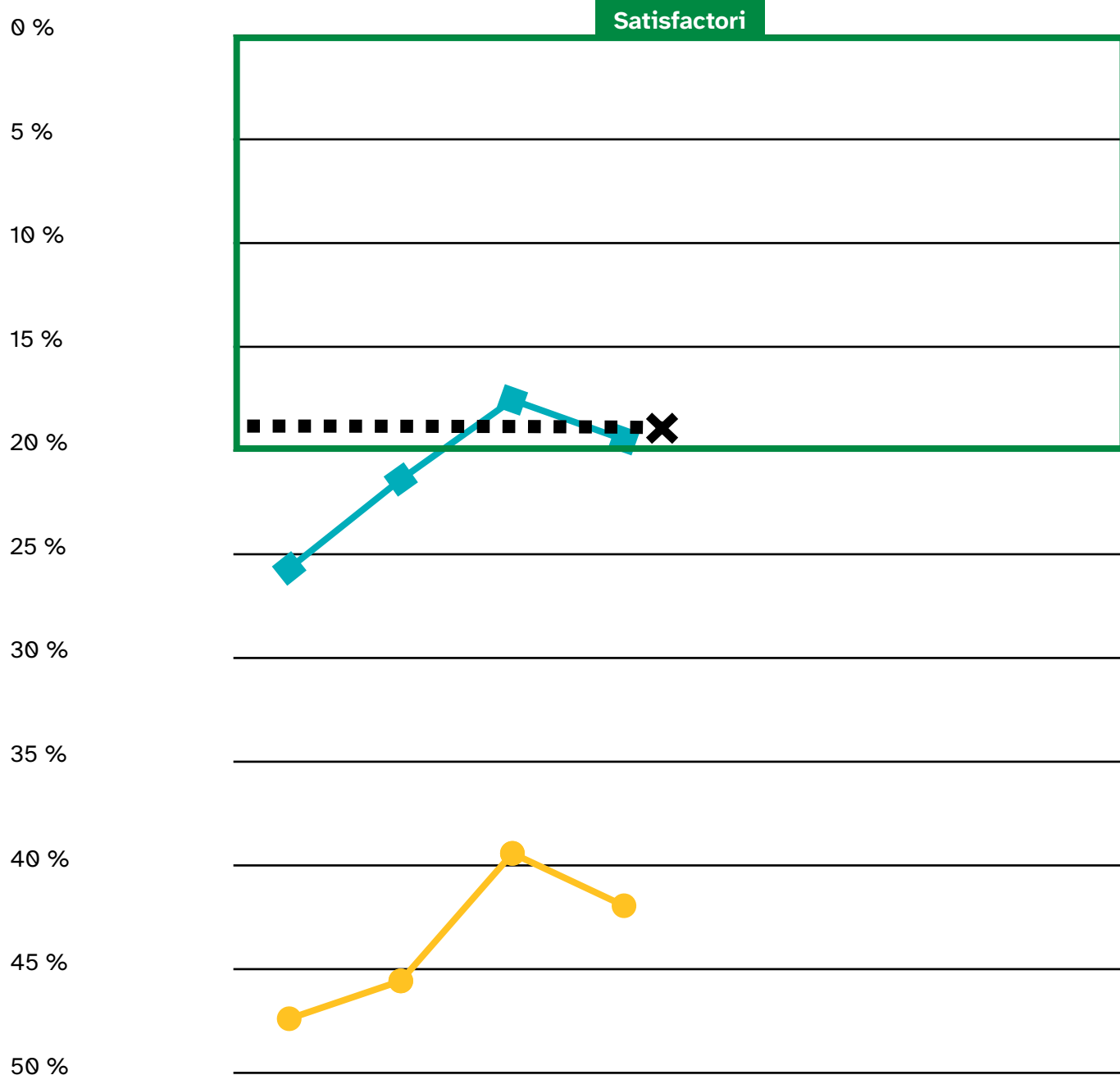
	Central amb interrupcions		Central amb canvi de llengua CA-ES				
● Google Meet	80,17 %	87,05 %	88,46 %	91,52 %	86,89 %	91,51 %	88,53 %
■ Microsoft Teams	89,87 %	92,33 %	92,09 %	95,26 %	90,70 %	94,49 %	91,90 %
✕ Mitjana del conjunt d'aplicacions	92,13 %						94,32 %

Figura 9. Puntuació WER per tipus de situació de comunicació no ideal.



	Central - preparat				Mala qualitat de so			
● Google Meet	37,30 %	35,63 %	41,71 %	35,29 %	41,62 %	38,51 %	49,14 %	38,50 %
■ Microsoft Teams	16,22 %	20,11 %	16,57 %	20,32 %	18,92 %	18,97 %	17,71 %	31,55 %
✕ Mitjana del conjunt d'aplicacions	17,35 %				18,75 %			

Font: Elaboració pròpia



	Entorn sorollós			
● Google Meet	47,31 %	45,40 %	39,46 %	41,95 %
■ Microsoft Teams	25,81 %	21,26 %	17,30 %	19,54 %
✕ Mitjana del conjunt d'aplicacions				18,85 %

Microsoft Teams, en canvi, presenta diferències significatives entre els discursos preparats —NER de 94,95 %—, els espontanis —NER 93,99 %— i, no tan lluny, els de condicions de producció no ideals —93,02 %—.

Com en el cas anterior, els millors resultats quant a varietat són en dialecte central —NER de discurs preparat de 96,59 %—, seguits pel nord-occidental —95,99 %—, el valencià —94,99 %— i, molt més allunyat, el balear —92,23 %—. Les puntuacions del discurs espontani són semblants, tot i que globalment més baixes.

En situacions de producció no ideals, els pitjors resultats són també en les mostres amb interrupcions dels parlants —NER més baixa en 4,35 % que els discursos espontanis del mateix dialecte en condicions ideals de producció—, seguits de les de canvi de llengua —NER més baixa en 2,56 %—, però els següents pitjors són en les de mala qualitat —WER més alta en 4,26 % que el mateix discurs en situació de producció ideal— i, finalment, les de soroll de fons —WER més alta en 2,67 %—.

Resultats específics

Si tenim en compte la quantitat i la gravetat dels errors en les situacions de producció ideals, observem que Google Meet és l'eina que en presenta més globalment. També és l'eina que en presenta més de tipus lleu i estàndard, i la segona que en presenta més de greus.

Per tipus, la gran majoria d'errors són de puntuació —el 37 % dels seus errors, que suposen el 42 % dels errors de puntuació de totes les aplicacions—, ja que les transcripcions d'aquesta eina prescindeixen de signes de puntuació i majúscules. Tot i que es pot considerar que generalment es tracta d'un error lleu, detectable i que no afecta la comprensió del significat, en alguns casos dona lloc a ambigüitats o falsos sentits.

Transcripció manual:

I res, amb moltes ganes. Estic estudiant ara màrqueting digital i a veure què tal.

Transcripció automàtica:

i res amb moltes **ganes estic** estudiant ara màrqueting digital i a veure què **tal**

Google Meet és l'eina que presenta més errors. La majoria són de puntuació, ja que les transcripcions d'aquesta eina prescindeixen de signes de puntuació i majúscules.

Transcripció manual:

[...] podia parlar més en aqueixes expressions i ara no, me costa més.

Transcripció automàtica:

[...] podia parlar més en expressions i ara **no me** costa més

En segon lloc, trobem un gran nombre d'errors de reconeixement sense causa aparent —el 27 %—, de totes les gravetats.

Transcripció manual:

Hi ha un home fora rentant el seu cotxe amb aigua i sabó.

Transcripció automàtica:

hi ha un home fora rentant el seu cotxe amb aigua i **segur**.

Transcripció manual:

[...] vam anar a l'illa de Borneo, que és on hi ha tots els orangutans, [...]

Transcripció automàtica:

[...] vam anar a **la línia** de borneo que és on hi ha tots els **hauran votants** [...]

Finalment, tot i que és l'eina que més errors ha comès en gairebé totes les categories —en puntuació i reconeixement, però també en omissions, errors per dubitació i reformulació, o canvi de temps verbals— cal destacar el gran nombre d'errors relacionats amb l'ortografia i les convencions lingüístiques —12 % dels seus errors, però el 31 % dels errors d'ortografia totals—.

Transcripció manual:

[...] em pose una pel·li i me'n vaig a dormir.

Transcripció automàtica:

[...] vam posar una **pele** me'n vaig a dormir.

Transcripció manual:

També està molt enfora d'on visc, l'aeroport, [...]

Transcripció automàtica:

també està molt **en fora** d'on vist aeroport [...]

A més, també val la pena mencionar el fet que aquesta eina sovint vacil·la a l'inici del discurs. En diverses mostres hem pogut observar que escriu i esborra algunes paraules abans de començar definitivament la transcripció.

Transcripció manual:

Quan era més petita anava de colònies i coses com fora, però aquest any no s'ha donat el cas. Llavors m'he passat com l'estiu treballant bastant.

Transcripció automàtica:

ni estiu treballant bastant

Transcripció manual:

Idò, i aqueix estiu jo he estat bastant per Mallorca, [...]

Transcripció automàtica:

bastant per **mallorca** [...]

Com a aspecte positiu, podem destacar que, juntament amb YouTube, és l'eina que menys ha estandarditzat les varietats dialectals. De la mateixa manera, aquesta eina tendeix a mantenir

en castellà els termes que reconeix correctament en les mostres amb barreja de llengües, tot i que en alguns casos en què la forma equivalent en català és molt semblant, la tradueix.

En el cas de Microsoft Teams, com passava amb les puntuacions generals, és la segona eina que presenta més errors globalment. A més, és l'eina que en presenta més de greus, i la segona que en presenta més de lleus i estàndard.

Per tipus d'error, la majoria són de reconeixement sense causa concreta —32 %—, seguits dels de puntuació —26 %—.

Transcripció manual:

És la col·lecció supervivent més gran de la seva classe.

Transcripció automàtica:

És la **connexió** supervivent més gran de la seva classe.

Transcripció manual:

[...] he après moltes coses, sobretot de com funcionen ses persones amb sordesa, [...]

Transcripció automàtica:

He après moltes coses, sobretot de com funcionen les persones **en surt deessa** [...]

Transcripció manual:

[...] veure que tu ets un referent per ells és una de les coses que a mi personalment em provoca més satisfacció.

Transcripció automàtica:

[...] veure que tu ets un referent per **ells. És** una de les coses que a mi personalment em provoca més satisfacció [...]

Transcripció manual:

[...] vam veure serps, vam veure peixos globus...

Transcripció automàtica:

[...] vam veure serps, vam veure **peixos, globus**, [...]

Tot i això, cal destacar que Microsoft Teams també ha comès un gran nombre d'errors d'ortografia i convencions lingüístiques —el 13 % dels seus errors, que suposen el 21 % dels errors d'ortografia de totes les aplicacions—. En aquest sentit, hem pogut observar que encara fa servir els accents diacrítics previs a l'actualització de la norma, i que substitueix la *ela* geminada (l·l) per dues *eles* separades per un espai (l l).

Transcripció manual:

I també viatjàvem molt pel... pel món, aprofitàvem tots els contactes que teníem [...]

Transcripció automàtica:

[...] i també viatjàvem molt per al món. **Aprofitavem** tots els contactes que teníem [...]

Transcripció manual:

Soc monitor des dels divuit anys [...]

Transcripció automàtica:

Sóc monitor des dels divuit anys [...]

Finalment, és l'aplicació que més errors ha comès a causa de la presència de castellanismes en els discursos —22 % dels errors d'aquest tipus de totes les aplicacions, sense tenir en compte les mostres de canvi de llengua—.

Microsoft Teams és la segona eina que presenta més errors.

Transcripció manual:

[...] també, o sea, sempre estem xerrant, [...]

Transcripció automàtica:

[...] també **ho serà** sempre estem xerrant [...]

Transcripció manual:

I bé, ara estàvem ahí fora, pues escollint els temes que anàvem a fer pels reportatges, pels perfils, per les entrevistes.

Transcripció automàtica:

i bé ara estaven fora, **però** escollint els temes que anàvem a fer pels reportatges, pels perfils, per les entrevistes.

En les mostres amb barreja de llengües, quan no ha comès errors, aquesta eina ha tendit a mantenir els termes en castellà, i només en alguns casos, en què la forma equivalent en català és molt semblant, els ha traduït.

Avaluació de les característiques d'accessibilitat

Si tenim en compte altres aspectes rellevants perquè la transcripció sigui accessible, podem destacar que les dues eines identifiquen els parlants, però ho fan a partir del micròfon d'on prové el so, de manera que si dues persones participen des d'un mateix ordinador, s'identificaran com un únic parlant. De fet, en analitzar les mostres amb interrupció dels parlants a partir d'una gravació, reproduïda des d'un únic ordinador, totes dues eines han comès

un gran nombre d'errors relacionats amb la barreja de discursos i intervencions de parlants diferents —Google Meet, el 39 % d'aquesta mena d'errors del conjunt de les aplicacions, i Microsoft Teams, el 40 %—.

Quant a la personalització, totes dues eines estan orientades a l'usuari i ofereixen la possibilitat d'ajustar la mida i el color de text, així com el color de fons. Google Meet també permet escollir la tipografia, i Microsoft Teams permet ajustar alguns aspectes de la posició dels subtítols.

Més particularment, en el cas de Google Meet destaca el fet que les transcripcions no incloguin cap mena de puntuació ni majúscules, ja que són elements que afecten significativament la comprensió del text, especialment per a aquelles persones amb sordesa que no puguin identificar clarament aspectes com el ritme o l'entonació del discurs. Aquest fet, això no obstant, té sentit si tenim en compte que el català està catalogat a l'aplicació com a llengua “beta”.

Pel que fa a Microsoft Teams, convé destacar que aquesta eina ofereix una primera versió de la transcripció en directe i, quan es produeix un canvi d'oració o una petita pausa, n'actualitza la puntuació i altres aspectes de convencions. Aquesta característica fa que l'usuari hagi de rellegir el text un cop acabada la intervenció per assegurar que ha interpretat el contingut correctament i, per tant, no seguir la conversa en directe. Així, no podem considerar que l'eina afavoreixi l'accessibilitat.

Speechmatics, AgilText i SCRIBAL poden ser un suport d'ús personal en espais de comunicació inaccessible.

Rendiment de les aplicacions de transcripció en directe i diferit (Speechmatics, AgilText i SCRIBAL)

Resultats globals

En termes generals, cap de les tres aplicacions supera globalment la puntuació mínima per considerar que presenten prou qualitat per ser eines d'accessibilitat que puguin complir amb l'obligació legal sobre accessibilitat, en substitució de les mesures de suport a la comunicació oral tradicionals.

Tot i això, totes tres s'hi apropen molt: Speechmatics presenta la millor puntuació del conjunt de les aplicacions —amb una mitjana WER de 22,68 % (acceptable) i NER de 97,58 % (3/10)—, seguit d'AgilText —WER 24,25 % (acceptable) i NER 97,49 % (3/10)— i SCRIBAL —WER 25,27 % (acceptable) i NER 96,89 % (2/10)—. En els tres casos, a més, diverses mostres superen individualment la puntuació mínima.

Per tant, sí que podem considerar que aquestes eines poden ser un suport d'ús personal per a les persones amb sordesa quan es troben

en espais de comunicació inaccessible o en certes situacions de comunicació.

Per tipus de mostra, Speechmatics presenta resultats significativament diferents entre els discursos preparats —mitjana NER de 98,94 %—, els espontanis —97,38 %— i els produïts en condicions no ideals —96,43 %—.

Pel que fa a les varietats dialectals, la millor puntuació és en el dialecte central, amb resultats molt alts —mitjana NER en discursos preparats de 99,25 %—, mentre que la resta de dialectes presenten resultats molt similars al voltant del 98,75 %. Els resultats dels discursos espontanis segueixen el mateix patró, tot i que amb puntuacions significativament més baixes, especialment en el dialecte balear, que es distancia per sota —mitjana NER de 96,50 %—.

Els resultats de Speechmatics en situacions de producció no ideals són diversos. D'una banda, observem que el soroll de fons i la mala qualitat no afecten gaire la qualitat de la transcripció —amb una mitjana WER respecte de les mostres equivalents en situació ideal més alta només en 0,57 % i 1,27 % respectivament—. D'altra banda, el canvi de llengua dona resultats significativament més baixos —NER més baixa en 1,63 %—, i la interrupció dels parlants, encara més —NER més baixa en 4,39 %—. Més concretament, en els textos amb interrupcions,

**AgilText és l'única eina
en què el canvi de llengua
no sembla suposar
un problema per al
reconeixement.**

aquesta eina ha comès errors tant pel que fa a no identificar els parlants com a barrejar o ometre les seves intervencions.

AgilText, per la seva banda, també presenta resultats significativament diferents entre discursos preparats —mitjana NER de 98,85 %—, espontanis —97,09 %— i produïts en condicions no ideals —96,53 %—.

Quant a les varietats dialectals, la millor puntuació és en els dialectes central i nord-occidental i també molt alta —mitjana NER en discursos preparats de 99,03 % i 98,95 %, respectivament— seguits molt de prop del valencià —98,83 %— i, encara prou a prop, el balear —98,60 %—. En el cas dels discursos espontanis, els resultats són significativament més baixos i s'incrementa la distància del central per amunt i del balear per baix.

Pel que fa a les situacions de producció no ideals, com en el cas anterior, el soroll de fons i la mala qualitat del so no afecten gaire el reconeixement —la mitjana WER respecte de les mostres equivalents és només 2,19 % més alta en el cas del soroll, i fins i tot 1,95 % més baixa en el cas de la mala qualitat—. De la mateixa manera, és l'única eina en què el canvi de llengua no sembla suposar un problema per al reconeixement —NER més baixa en només 0,56 % respecte als textos equivalents—, cosa que resulta especialment positiva si tenim en compte que la barreja de català i castellà és un fenomen freqüent dins les situacions de comunicació reals del nostre entorn sociolingüístic.

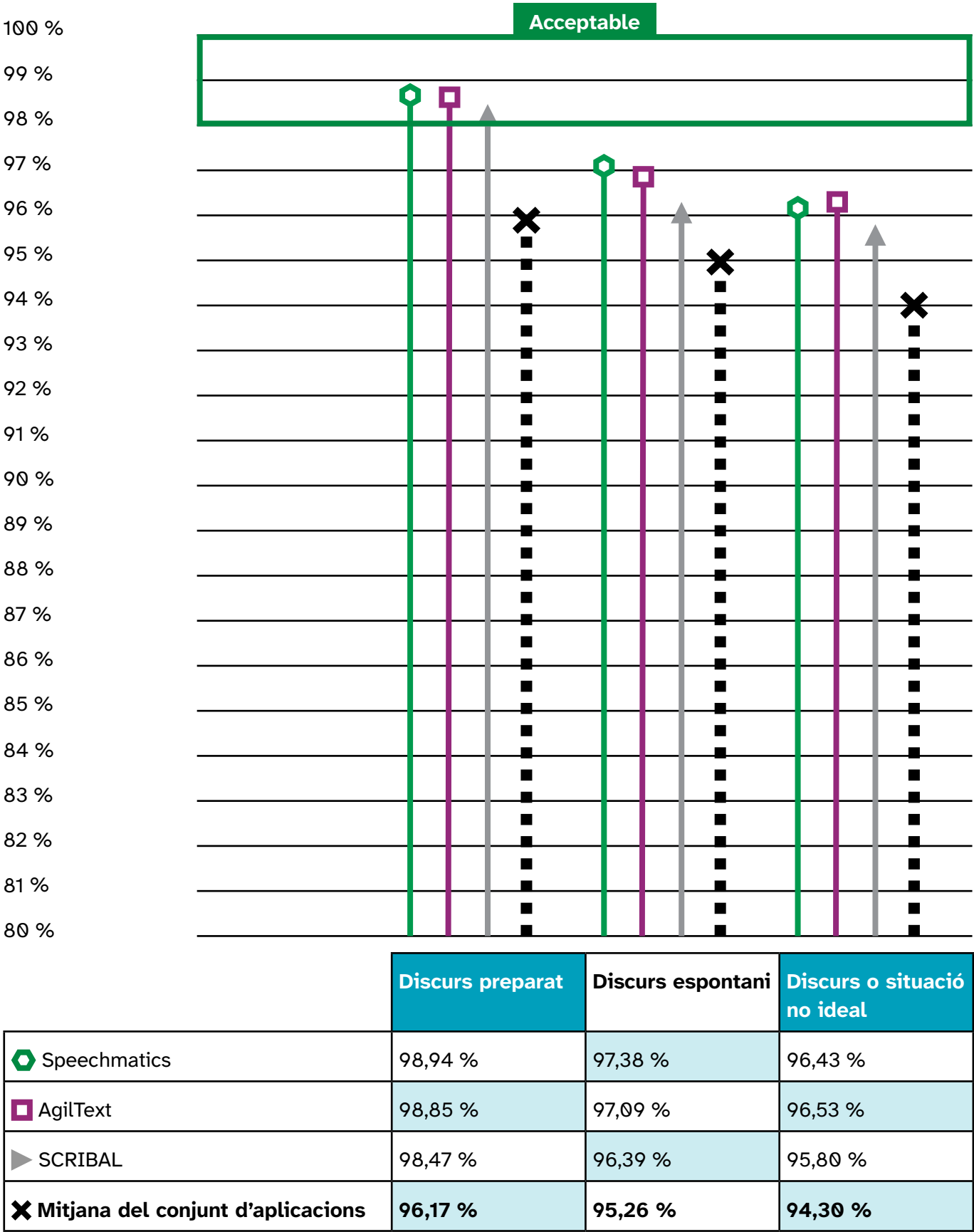
La interrupció dels parlants, en canvi, sí que dona resultats molt més baixos —NER més baixa en 3,10 %—. En aquest cas, tot i que AgilText comet alguns errors en la identificació dels parlants, la majoria dels casos són d'omissió d'intervencions, tot i que sovint són per confirmar o reafirmar el discurs de l'altra persona, de manera que es tracta d'errors lleus.

Quant a SCRIBAL, també presenta resultats significativament diferents entre discursos preparats —mitjana NER de 98,47 %—, espontanis —96,39 %— i produïts en condicions no ideals —95,80 %—.

Pel que fa a les varietats dialectals, les millors puntuacions són també altes, en els dialectes nord-occidental i central —mitjana NER en discursos preparats de 98,96 % i 98,79 %, respectivament— seguides del valencià —98,40 %— i, una mica més lluny, el balear —97,73 %—. En el cas dels discursos espontanis, els resultats són significativament més baixos i s'incrementa la distància del nord-occidental per amunt, seguit del valencià i deixant enrere el central i el balear, amb puntuacions semblants. No obstant això, cal mencionar que diverses de les mostres de discurs espontani de SCRIBAL compten amb errors greus per l'omissió de fragments, que no tenen a veure amb la varietat dialectal —s'explica en detall més endavant—, de manera que els resultats no són del tot representatius de la influència d'aquesta variable en el reconeixement.

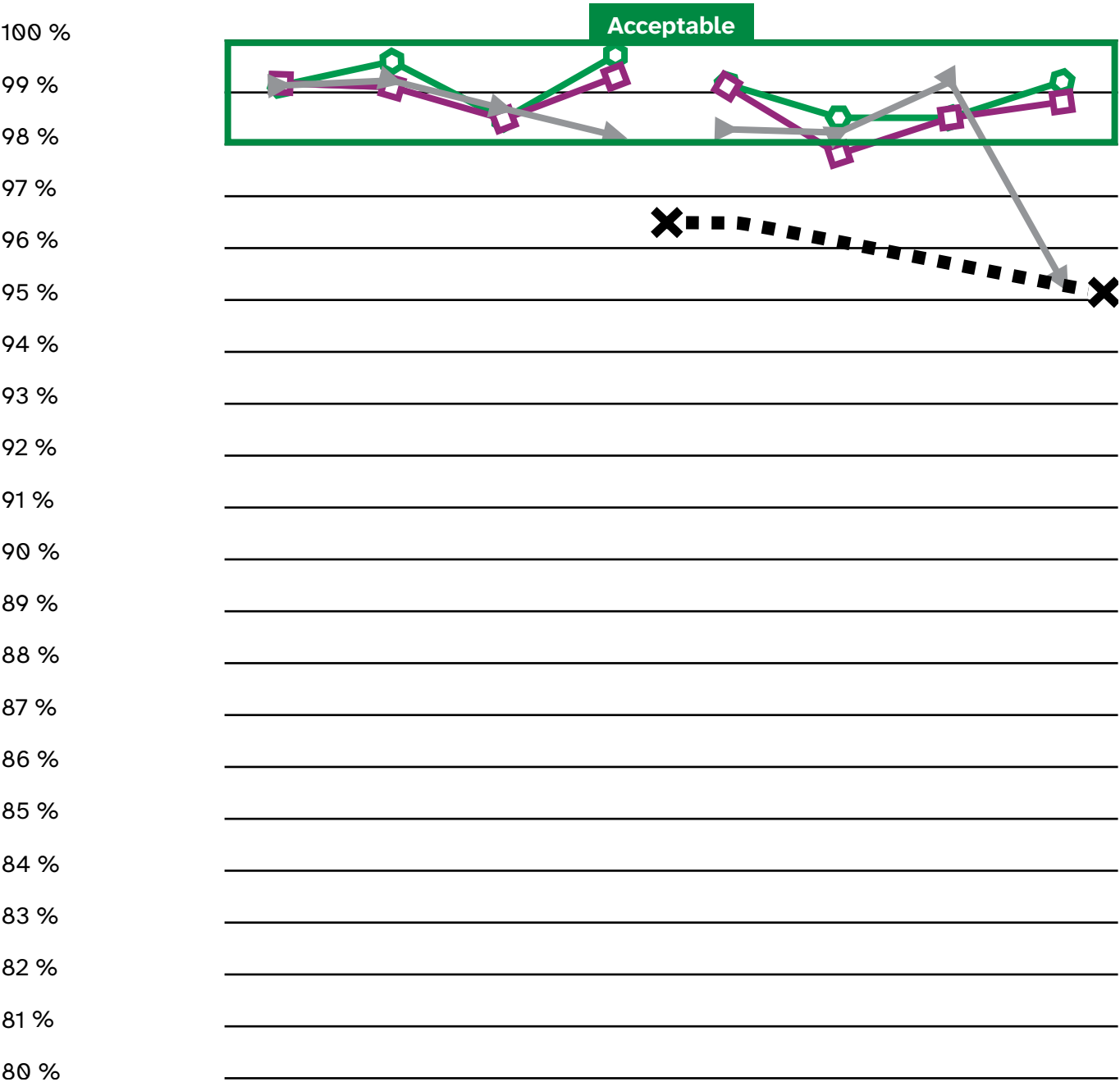
Pel que fa a les situacions de producció no ideals, a diferència dels casos anteriors, el soroll de fons i la mala qualitat del so sí que afecten el reconeixement —la mitjana WER respecte de les mostres equivalents és 7,70 % més alta en el cas del soroll, i 3,60 % més alta en el cas de la mala qualitat—. Quant al canvi de llengua i la interrupció dels parlants, no podem considerar que els resultats són representatius, ja que els textos de discurs espontani de referència contenen omissions de grans fragments —s'analitza en detall als resultats específics—. En el cas de les interrupcions, però, destaca el fet que no ha comès cap error de barreja d'intervencions de parlants diferents, tot i que mai els identifica. La majoria d'errors, per tant, són d'omissió d'intervencions dels parlants.

Figura 10. Puntuació NER per tipus de mostra.



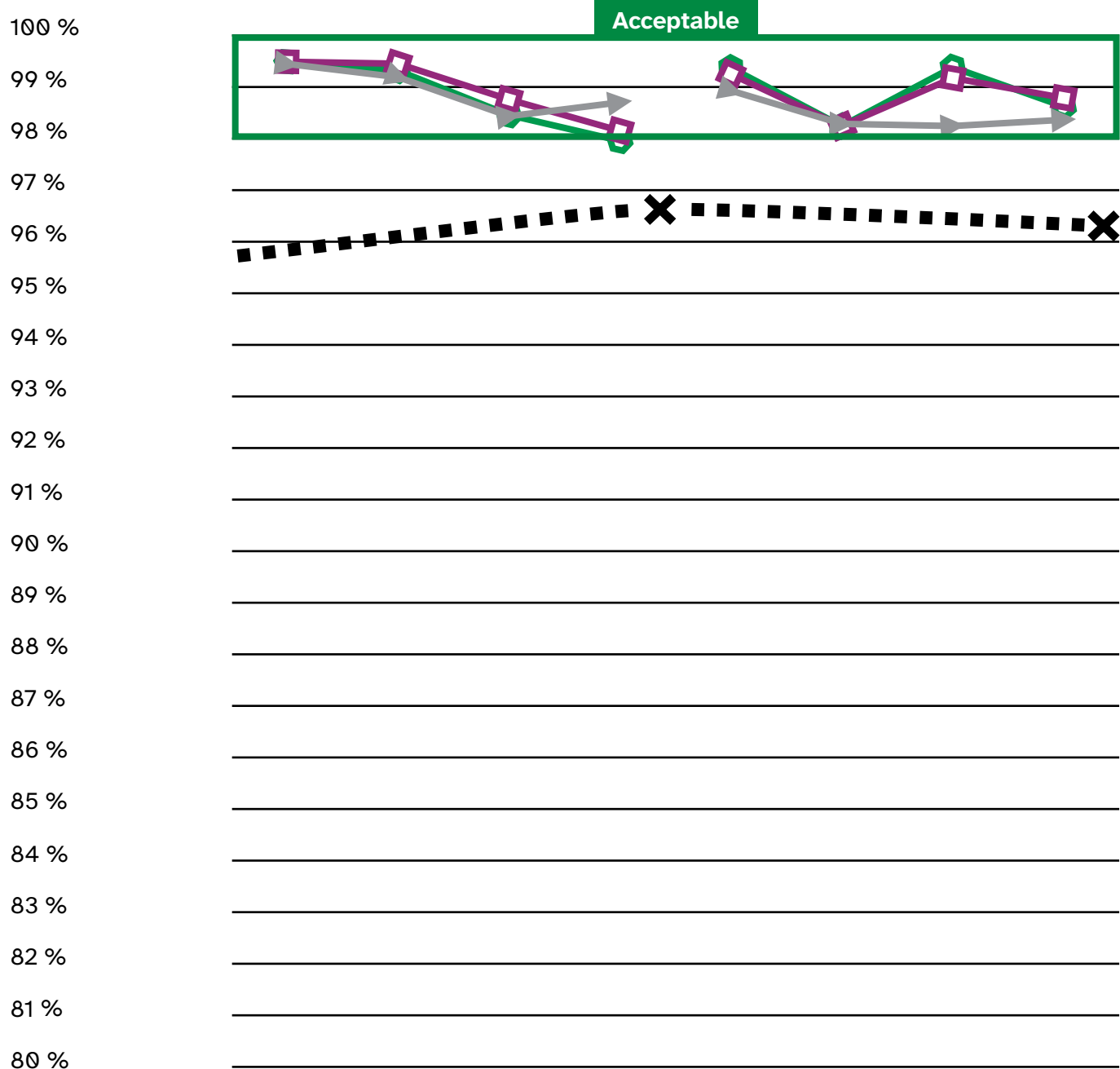
Font: Elaboració pròpia

Figura 11. Puntuació NER dels discursos preparats per varietat dialectal.



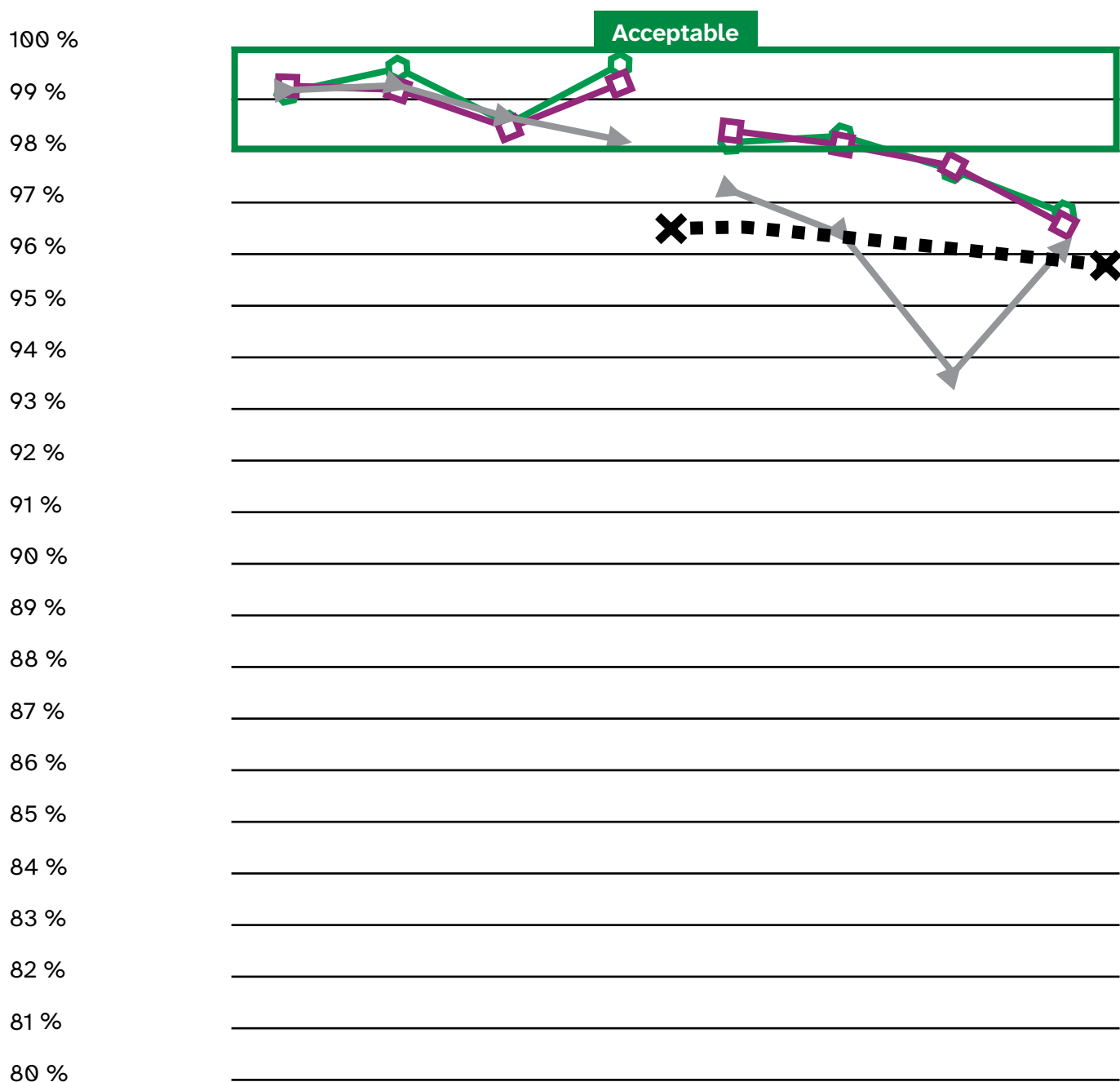
	Central				Balear			
Speechmatics	99,04 %	99,62 %	98,59 %	99,76 %	99,12 %	98,52 %	98,54 %	99,12 %
AgilText	99,20 %	99,12 %	98,49 %	99,31 %	99,13 %	97,87 %	98,53 %	98,87 %
SCRIBAL	99,07 %	99,20 %	98,74 %	98,16 %	98,25 %	98,20 %	99,16 %	95,30 %
Mitjana del conjunt d'aplicacions	96,54 %				95,11 %			

Font: Elaboració pròpia



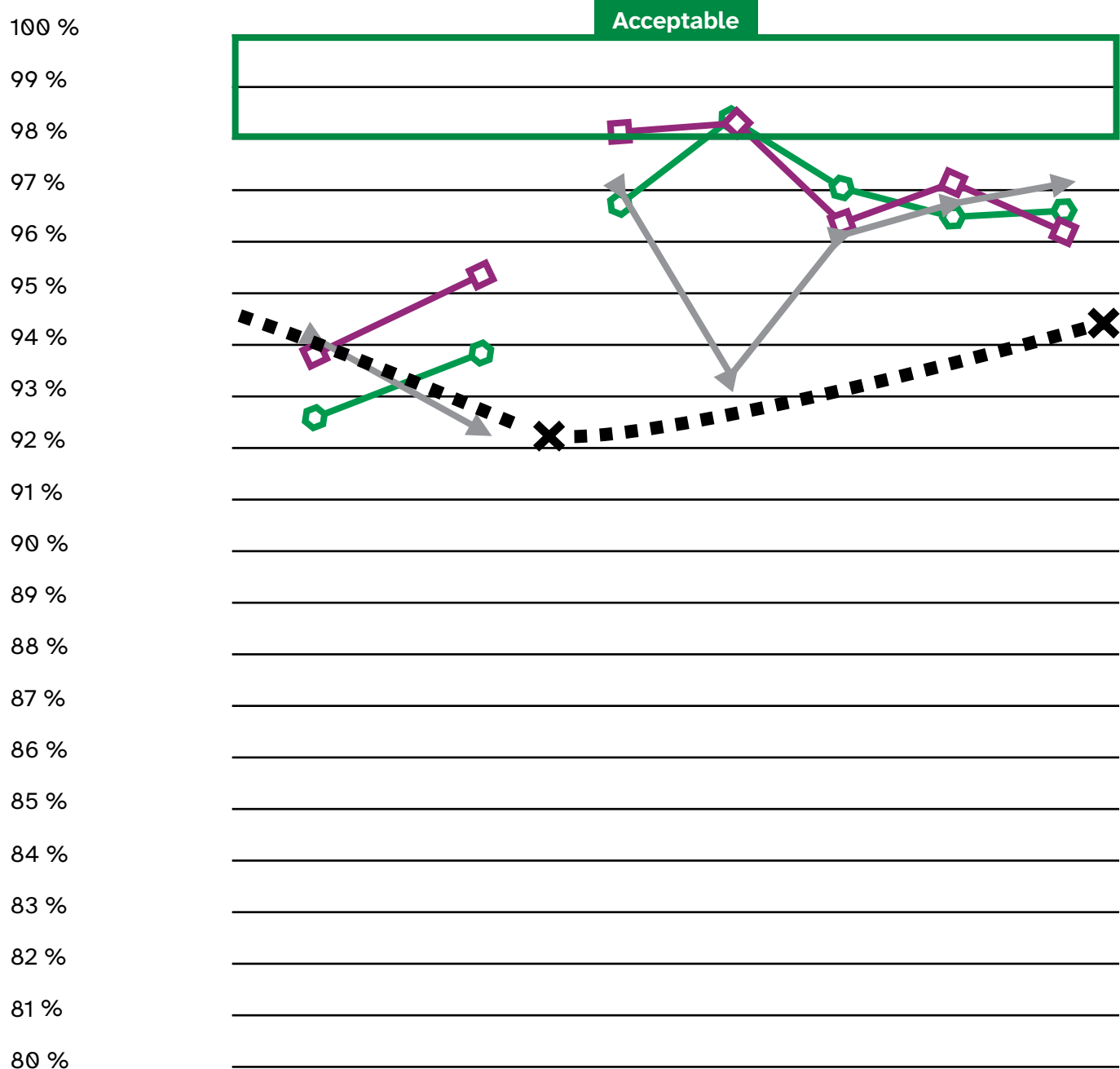
	Nord-occidental				València			
Speechmatics	99,49 %	99,28 %	98,41 %	97,98 %	99,36 %	98,18 %	99,36 %	98,67 %
AgilText	99,49 %	99,42 %	98,77 %	98,10 %	99,24 %	98,15 %	99,11 %	98,80 %
SCRIBAL	99,49 %	99,12 %	98,51 %	98,72 %	98,95 %	98,17 %	98,16 %	98,33 %
✕ Mitjana del conjunt d'aplicacions	96,74 %				96,30 %			

Figura 12. Puntuació NER per tipus de discurs.



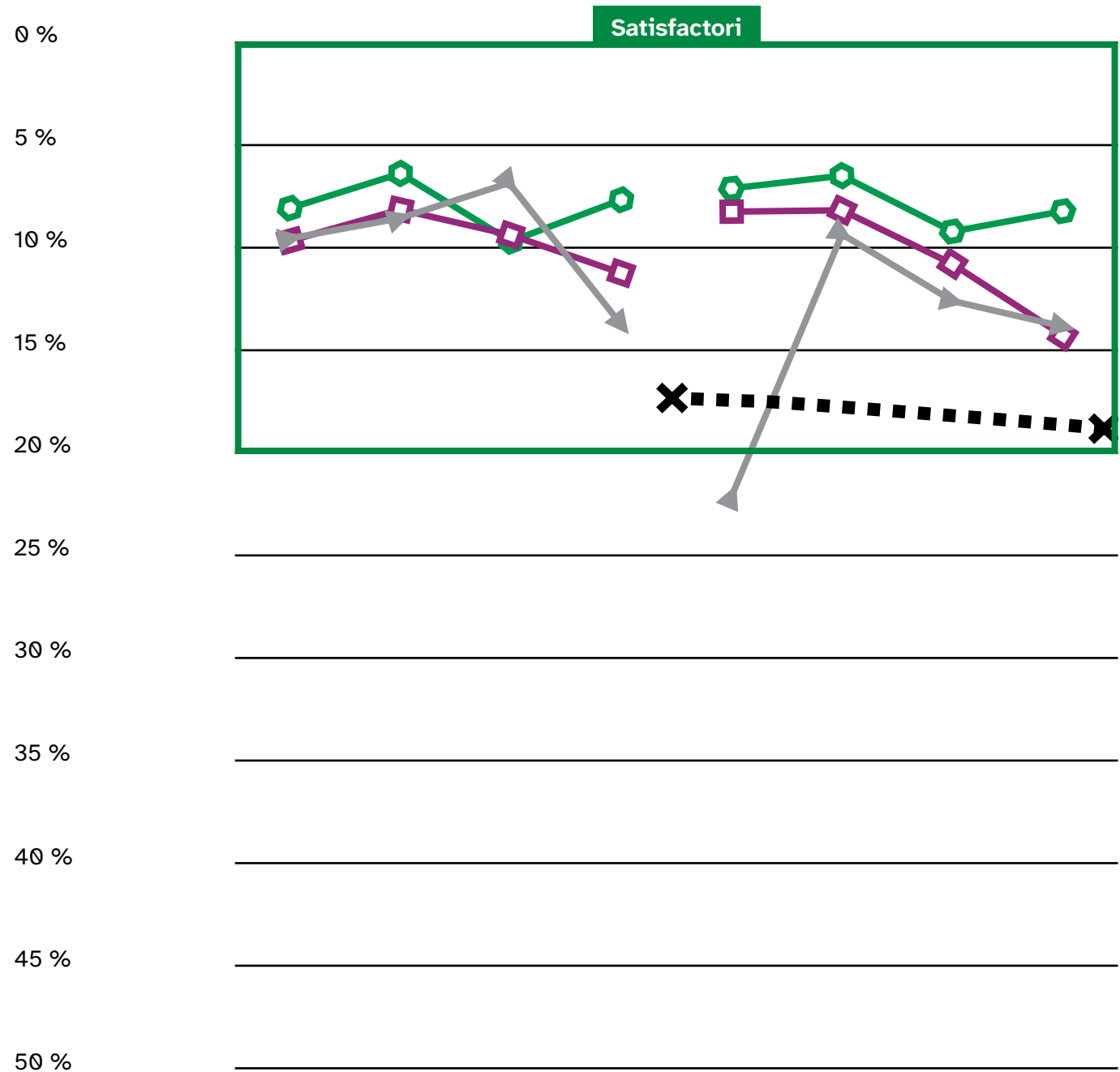
	Central - preparat				Central - espontani			
Speechmatics	99,04 %	99,62 %	98,59 %	99,76 %	98,08 %	98,22 %	97,64 %	96,90 %
AgilText	99,20 %	99,12 %	98,49 %	99,31 %	98,40 %	98,06 %	97,82 %	96,62 %
SCRIBAL	99,07 %	99,20 %	98,74 %	98,16 %	97,14 %	96,30 %	93,82 %	96,09 %
X Mitjana del conjunt d'aplicacions	96,54 %				95,91 %			

Font: Elaboració pròpia



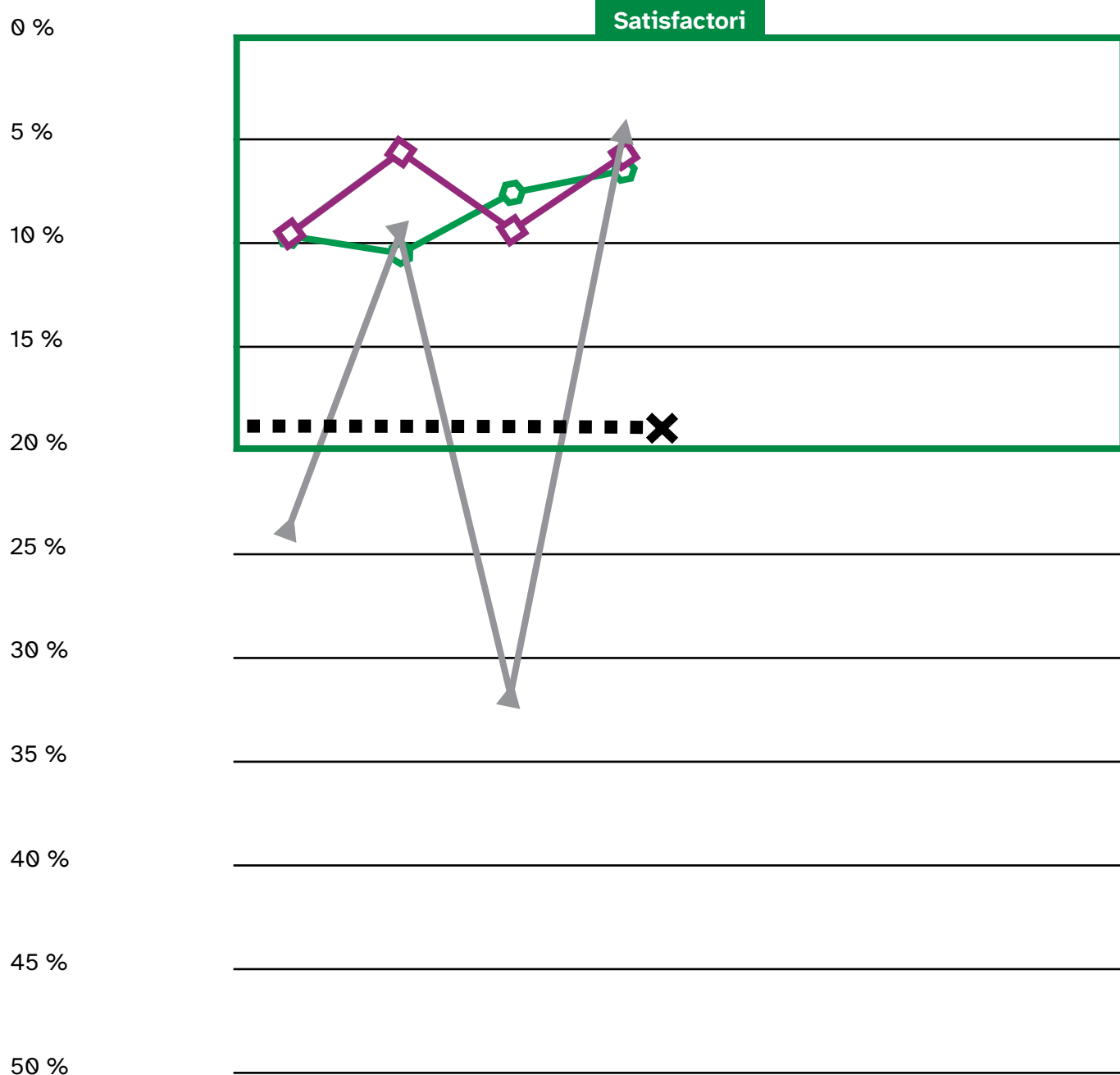
	Central amb interrupcions		Central amb canvi de llengua CA-ES				
Speechmatics	92,71 %	93,92 %	96,88 %	98,38 %	97,02 %	96,46 %	96,68 %
AgilText	93,94 %	95,31 %	98,04 %	98,27 %	96,35 %	97,05 %	96,13 %
SCRIBAL	94,06 %	92,26 %	97,00 %	93,42 %	96,05 %	96,78 %	97,04 %
✕ Mitjana del conjunt d'aplicacions		92,13 %					94,32 %

Figura 13. Puntuació WER per tipus de situació de comunicació no ideal.



	Central - preparat				Mala qualitat de so			
Speechmatics	8,11 %	6,32 %	9,71 %	7,49 %	7,03 %	6,32 %	9,14 %	8,02 %
AgilText	9,73 %	8,05 %	9,14 %	11,23 %	8,11 %	8,05 %	10,86 %	14,44 %
SCRIBAL	9,73 %	8,62 %	6,86 %	13,90 %	22,16 %	9,20 %	12,57 %	13,90 %
Mitjana del conjunt d'aplicacions	17,35 %				18,75 %			

Font: Elaboració pròpia



	Entorn sorollós			
Speechmatics	9,68 %	10,34 %	7,57 %	6,32 %
AgilText	9,68 %	5,75 %	9,19 %	5,75 %
SCRIBAL	23,66 %	9,77 %	31,89 %	4,60 %
✕ Mitjana del conjunt d'aplicacions				18,85 %

Resultats específics

Si tenim en compte la quantitat i la gravetat dels errors en les situacions de producció ideals, observem que Speechmatics és l'eina que menys en presenta globalment, i que menys en presenta de tipus greu i estàndard.

Per tipus, la majoria d'errors són de puntuació —28 %—, seguits dels de reconeixement sense causa aparent —17 %—, tot i que globalment és l'eina que menys n'ha comès d'aquest darrer tipus —5 % de tots els errors de reconeixement sense causa—.

Transcripció manual:

No, espera, aquest cap de setmana tenim algo més.

Transcripció automàtica:

No espera Aquest cap de setmana tenim algú més.

Transcripció manual:

Després es pròxim llibre que llegiré no sé quin serà.

Transcripció automàtica:

Després el pròxim llibre que **llegiré. No** sé quin serà.

Transcripció manual:

[...] que passen més temps quasi amb tu com a profe [...]

Transcripció automàtica:

[...] que passen més temps **aquí** amb tu com a profe, [...]

Transcripció manual: [...]

al setembre, que ha sigut quan no treballava ni feia res, [...]

Transcripció automàtica:

[...] al setembre, que ha sigut quan no treballava **en FR.**

Pel que fa a la resta de tipus d'errors, els resultats estan molt equilibrats. Això no obstant, destaca en ser l'aplicació que més ha estandarditzat les mostres —amb el 21 % de canvis dialectals del conjunt de les aplicacions—, apropant totes les varietats a l'estàndard o, en alguns casos, al central.

Speechmatics és l'eina que menys errors presenta.

Speechmatics és l'eina que més correccions lingüístiques ha fet.

Transcripció manual:

Jo toque... Toque el trombó.

Transcripció automàtica:

Jo **toco, toco** el trombó.

Transcripció manual:

[...] i sí que és vera que es transport és una liada, [...]

Transcripció automàtica:

I sí que és **cert** que el transport és una aliada, [...]

Transcripció manual:

[...] aquest any me tocarà pagar més.

Transcripció automàtica:

[...] aquest any **em** tocarà pagar més.

En aquest sentit, Speechmatics també tendeix, en les mostres amb barreja de llengua, a traduir els termes en castellà al català.

Finalment, cal mencionar també que és l'eina que més correccions lingüístiques ha fet, especialment quant a l'eliminació de dubitacions i reformulacions.

Transcripció manual:

Bueno, són... són alumnes que venen cada dia a classe, [...]

Transcripció automàtica:

Bé, **són** alumnes que vénen cada dia a classe, [...]

Com a altres aspectes positius, podem destacar que és l'eina que menys errors ha fet relacionats tant amb el canvi de temps verbals com amb l'ús de noms propis.

En el cas d'AgilText, és la segona eina que menys errors presenta globalment, i també la segona que menys en presenta de tipus greu i estàndard.

Per tipus, la gran majoria d'errors són de puntuació —31 %—, seguits dels de reconeixement sense causa aparent —19 %—.

Transcripció manual:

[...] amb es màster, com l'estudii a distància puc anar més a Menorca [...]

Transcripció automàtica:

[...] amb el màster, com l'estudia a **distància?**
Puc anar més a Menorca [...]

AgilText és la segona eina que menys errors presenta.

Transcripció manual:

[...] quan arribo allà m'he de creuar tot Dinamarca [...]

Transcripció automàtica:

[...] quan arribo **allà. M'he** de creuar tot Dinamarca [...]

Transcripció manual:

[...] perquè li vaig regalar pel seu aniversari [...]

Transcripció automàtica:

[...] perquè **la** vaig regalar pel seu aniversari [...]

Transcripció manual:

A l'exèrcit britànic, cada regiment i cos té un distintiu diferent.

Transcripció automàtica:

[...] a l'exèrcit britànic, **que ha** regiment i cos té un distintiu diferent.

Pel que fa a la resta de tipus d'errors, com en el cas anterior, els resultats estan molt equilibrats i només destaca una tendència a estandarditzar les mostres —amb el 19 % de canvis dialectals del conjunt de les aplicacions—.

Transcripció manual:

[...] i açò també me fa molta il·lusió.

Transcripció automàtica:

[...] i **això** també **em** fa molta il·lusió.

Transcripció manual:

A Alcalà de Xivert hi anam el cinc d'agost.

Transcripció automàtica:

Alcalà de Xivert. Hi **anem** el cinc d'agost.

Com en el cas anterior, AgilText també tendeix, en les mostres amb barreja de llengua, a traduir els termes en castellà al català.

Finalment, cal mencionar també que és la tercera eina que més correccions lingüístiques ha fet, especialment quant a l'eliminació de dubitacions i reformulacions, però també a la correcció d'errors gramaticals i castellanismes.

Transcripció manual:

O sea, va marxar de... de casa.

Transcripció automàtica:

O sigui, va marxar **de** casa, [...]

Com a altres aspectes positius, convé destacar que és l'eina que menys addicions o al·lucinacions ha presentat i, sobretot, que és

AgilText és l'eina que menys addicions o al·lucinacions ha presentat i l'única que no ha comès cap error en l'ús de xifres.

l'única que no ha comès cap error en l'ús de xifres, tot i que sí que n'ha comès relacionats amb les convencions d'aquestes xifres, com escriure-les amb majúscula inicial o barrejar l'escriptura amb paraules i amb símbols. Aquest fet és rellevant, però cal tenir en compte que els casos de presència de xifres en els discursos de la nostra mostra eren reduïts.

Quant a SCRIBAL, es troba en un punt mitjà de la taula pel que fa a nombre d'errors, tot i que més propera a Speechmatics i a AgilText que a la resta. En aquest sentit, cal destacar que és la segona aplicació que menys errors lleus ha comès.

Per tipus, la gran majoria d'errors són de reconeixement sense causa aparent —25 %—, seguits, a diferència de la majoria d'eines, de les omissions —17 %—. En aquest cas, però, cal mencionar que aquestes omissions són de grans fragments de text i, tenint en compte que l'eina encara està en fase de prova, no podem determinar si es tracta d'errors de reconeixement de veu o d'altres aspectes informàtics. Si ometem aquesta mena d'errors, els següents més freqüents són de puntuació —14 %—.

Transcripció manual:

I després al setembre vaig de càmping amb les meves amigues, [...]

Transcripció automàtica:

I després, al setembre, em vaig de **campionat** amb les meves amigues.

Transcripció manual:

[...] hem anat com d'excursions a gorgs i basses i coses aixins.

Transcripció automàtica:

Hem anat com d'excursions a **corcs** i basses i coses així.

Transcripció manual:

La veritat és que donen pluges per a aquesta setmana, cosa que tampoc és guai, [...]

Transcripció automàtica:

[...] la veritat és que vénen pluges per aquesta **setmana. Cosa** que tampoc és guai, [...]

Transcripció manual:

[...] i per això he hagut de fer entrevistes per videollamada, [...]

Transcripció automàtica:

[...] i per això ho he hagut de **fer. Entrevistes** per videollamada [...]

Pel que fa a la resta de tipus d'errors, com en el cas anterior, els resultats estan molt equilibrats, i només destaca lleugerament una tendència a cometre errors quan hi ha repeticions o reformulacions en els discursos espontanis —15 % dels errors—.

El segon error més freqüent de SCRIBAL són les omissions de grans fragments de text. L'eina encara està en fase de prova i no podem determinar si són errors de reconeixement o informàtics.

Transcripció manual:

I ja no... Fa uns anys me venien més al cap, [...]

Transcripció automàtica:

Ja no, amb fa uns anys me venien més al cap.

Com a aspecte positiu, podem destacar que és l'eina que menys errors d'ortografia i ús de majúscules ha comès.

Avaluació de les característiques d'accessibilitat

Si tenim en compte altres aspectes rellevants perquè la transcripció sigui accessible, podem destacar el fet que tant Speechmatics com AgilText identifiquen els parlants, tot i que en totes dues aplicacions cal activar-ne l'opció prèviament. En el primer cas, la identificació es fa a través d'una etiqueta —“Parlant 1”, “Parlant 2”, etc.—, mentre que en el segon cas és a través d'un guió llarg a l'inici de la primera intervenció quan el parlant canvia. En el cas de SCRIBAL, no s'identifiquen.

Quant a la personalització, trobem en AgilText una eina orientada a tècnics d'accessibilitat i usuaris, amb una interfície amable i opcions

ajustables pel que fa a la visualització. Més concretament, és l'única eina d'aquest grup que permet ajustar aspectes com la mida, la tipografia i el color de lletra, així com l'interlineat i el color de fons. A més, també permet veure els subtítols a la seva interfície, en pantalla completa o com a una nova finestra del navegador —ajustable en mida i forma—. D'altra banda, també permet personalitzar la transcripció incloent-hi un glossari propi.

En el cas de Speechmatics, l'eina està orientada a un perfil més especialitzat en l'àmbit del reconeixement de veu, amb una interfície menys amable i opcions ajustables només pel que fa a característiques tècniques de la transcripció —ús de certs signes de puntuació, rapidesa en l'aparició de subtítols o detecció d'elements que no són de parla—, tot i que també permet el mode de pantalla completa. En aquest estudi, les transcripcions analitzades s'han obtingut amb la configuració bàsica.

AgilText és una eina orientada a tècnics d'accessibilitat i usuaris, amb una interfície amable i opcions ajustables pel que fa a la visualització, mentre que Speechmatics està orientada a un perfil especialitzat en l'àmbit del reconeixement de veu.

Rendiment de les aplicacions de transcripció de continguts en diferit (YouTube i Softcatalà)

Resultats globals

En termes generals, cap de les dues aplicacions supera globalment la puntuació mínima per considerar que presenten prou qualitat per ser eines d'accessibilitat per a persones amb sordesa que donin resposta a l'obligació legal. Tot i això, Softcatalà presenta una mitjana WER de 26,24 % (acceptable) i NER de 96,56 % (1/10), prou per sobre de YouTube, que, tot i tenir una mitjana WER de 27,75 % (acceptable), presenta una NER de 94,47 % (0/10).

Per tant, Softcatalà pot ser una eina útil de suport a l'accessibilitat —en certes situacions i contextos, i sempre que no existeixi una subtitulació professional de qualitat— per accedir al contingut de productes audiovisuals. No podem recomanar, en canvi, fer servir YouTube per a aquesta finalitat.

Per tipus de mostra, Softcatalà presenta resultats molt similars en els discursos preparats —amb una mitjana NER de 96,76 %—, espontanis —NER 96,61 %— i en condicions de producció no ideals —NER 96,32 %—, tot i que segueix la tendència comuna.

De la mateixa manera, tampoc no presenta diferències significatives entre varietats dialectals. Els millors resultats són els de català central —mitjana NER de 96,98 % en discursos preparats—, seguits del nord-occidental —96,95 %—, el balear —96,43 %— i el valencià —96,46 %—. Els resultats dels discursos espontanis són molt similars.

Quant a les situacions de producció no ideals, el soroll de fons no sembla afectar la qualitat de la transcripció, ja que les mostres presenten una mitjana WER en discursos preparats més baixa

en només 0,21 %. Sí que donen pitjors resultats, tot i que de manera poc significativa, les mostres de mala qualitat —WER de discurs preparat més baix en 4 %—, les de canvi de llengua —NER més baixa en 1,13 %— i les d'interrupcions —NER més baixa en 2,13 %—. En aquest sentit, destaca el fet que Softcatalà no comet cap error de barreja d'intervencions de parlants diferents, tot i no identificar-los mai. Sí que comet, però, omissions d'intervencions, tot i que en la majoria dels casos es tracta de confirmacions de què diu l'altra persona i són errors lleus.

D'altra banda, YouTube és l'única aplicació que presenta millors resultats als discursos espontanis —amb una mitjana NER de 95,25 %— que als preparats —NER 94,82 %—, tot i que sí que trobem la pitjor puntuació a les mostres en condicions de producció no ideals —NER 93,33 %—.

Al contrari que l'aplicació anterior, YouTube presenta molta heterogeneïtat en els resultats. Per varietats dialectals, el nord-occidental i el valencià obtenen puntuacions millors i més estables —NER de discursos preparats de 96,20 % i 95,75 %, respectivament—, mentre que el central i el balear presenten puntuacions més baixes i disperses entre mostres concretes —NER de 93,12 % i 94,21 %, respectivament—. En els discursos espontanis, les primeres varietats es mantenen més o menys en les mateixes puntuacions, així com el balear. El central, en canvi, dona molts millors resultats —95,71 %—.

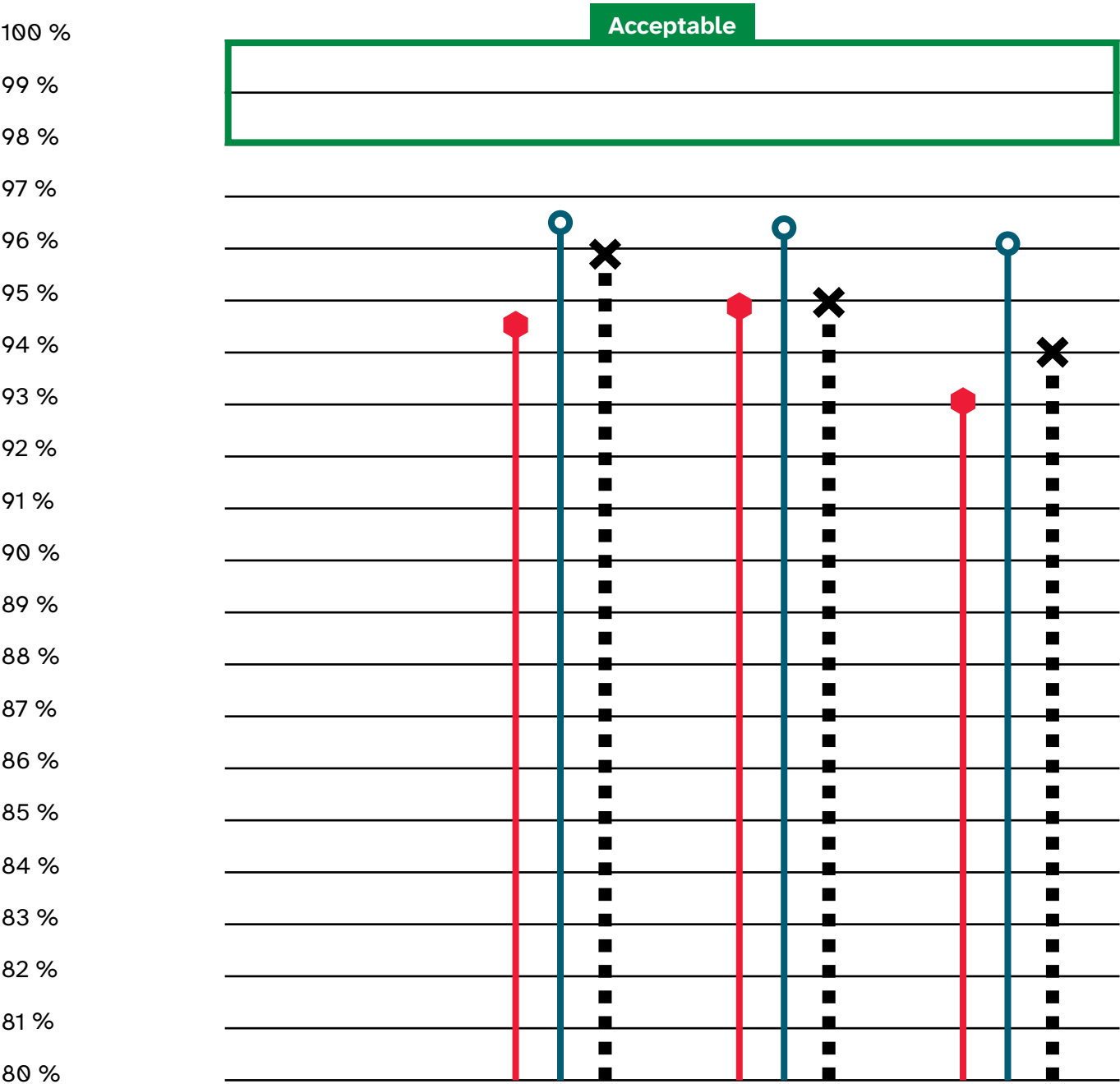
Quant a les situacions de producció no ideals, el soroll de fons i la mala qualitat no semblen afectar la qualitat de les transcripcions de YouTube. De fet, els resultats són lleugerament millors en aquestes mostres que en les mostres equivalents en condicions ideals —WER més baixa en 4,72 % amb soroll de fons i en 4 % amb mala qualitat—. Sí que donen resultats lleugerament pitjors les mostres amb interrupció —NER més baixa en 1,82 %— i, en el cas del canvi de llengua, significativament pitjors —NER més alta en 4,3 %—.

**Softcatalà pot ser una
eina útil de suport a
l'accessibilitat per accedir
al contingut de productes
audiovisuals.**

YouTube presenta molta heterogeneïtat en els resultats.

Així, els resultats de YouTube són gairebé contraris a les tendències de la resta d'eines i al que seria més intuïtiu, i destaca el fet que obtingui els millors resultats en situacions de comunicació no ideals, especialment amb soroll de fons, mala qualitat i discursos no preparats. No obstant això, desconeixem si aquest fet es deu a un entrenament específic per a aquestes variables —tenint en compte el tipus de vídeos de la plataforma— o si es pot deure a un biaix provocat en convertir els enregistraments d'àudio a un format de vídeo.

Figura 14. Puntuació NER per tipus de mostra.



	Discurs preparat	Discurs espontani	Discurs o situació no ideal
YouTube	94,82 %	95,25 %	93,33 %
Softcatalà	96,76 %	96,61 %	96,32 %
Mitjana del conjunt d'aplicacions	96,17 %	95,26 %	94,30 %

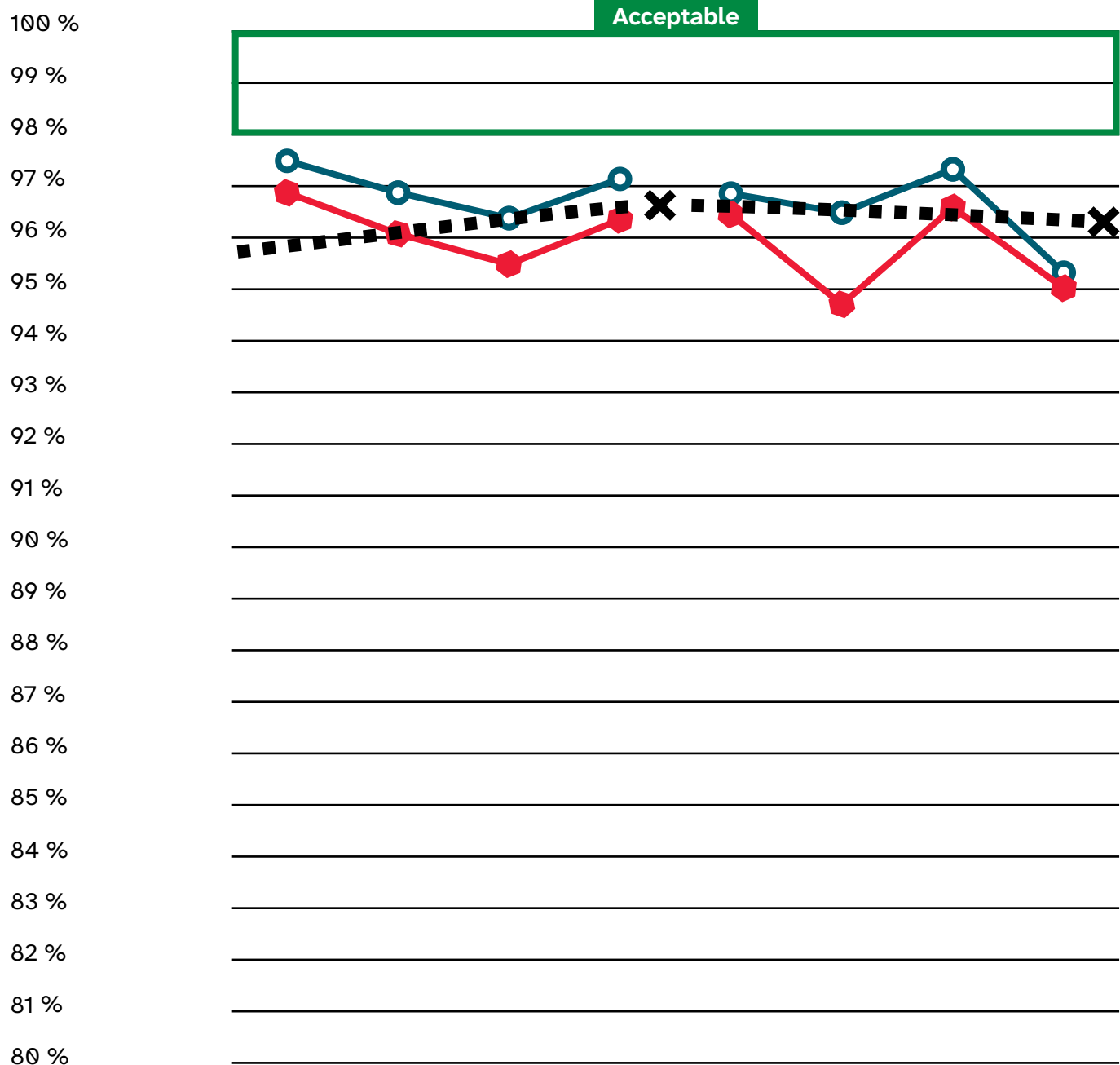
Font: Elaboració pròpia

Figura 15. Puntuació NER dels discursos preparats per varietat dialectal.



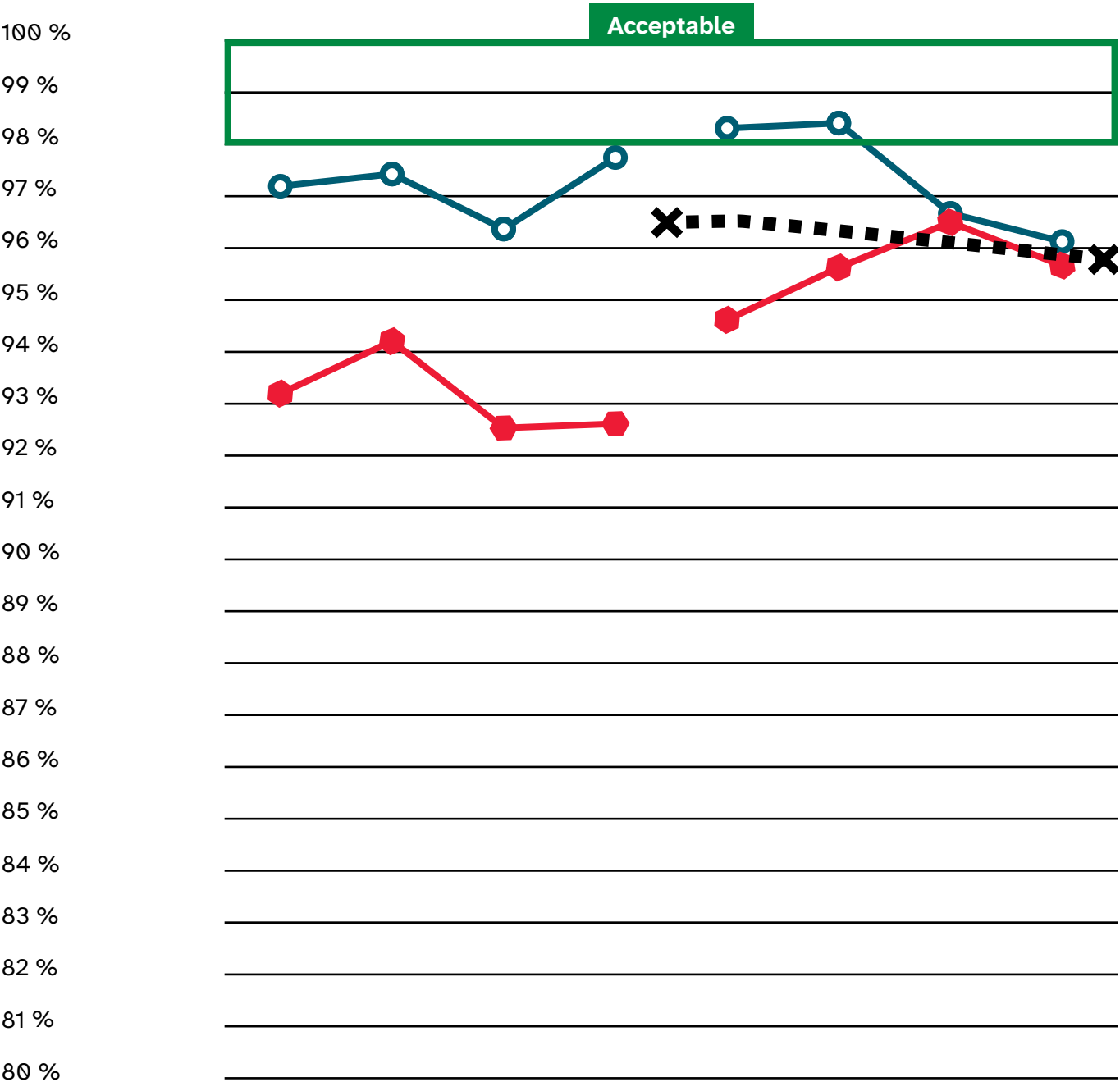
	Central				Balear			
● Softcatalà	97,09 %	97,45 %	96,39 %	97,94 %	96,27 %	95,77 %	98,18 %	95,49 %
● YouTube	93,13 %	94,08 %	92,59 %	92,68 %	94,04 %	93,72 %	94,72 %	94,34 %
✕ Mitjana del conjunt d'aplicacions	96,54 %				95,11 %			

Font: Elaboració pròpia



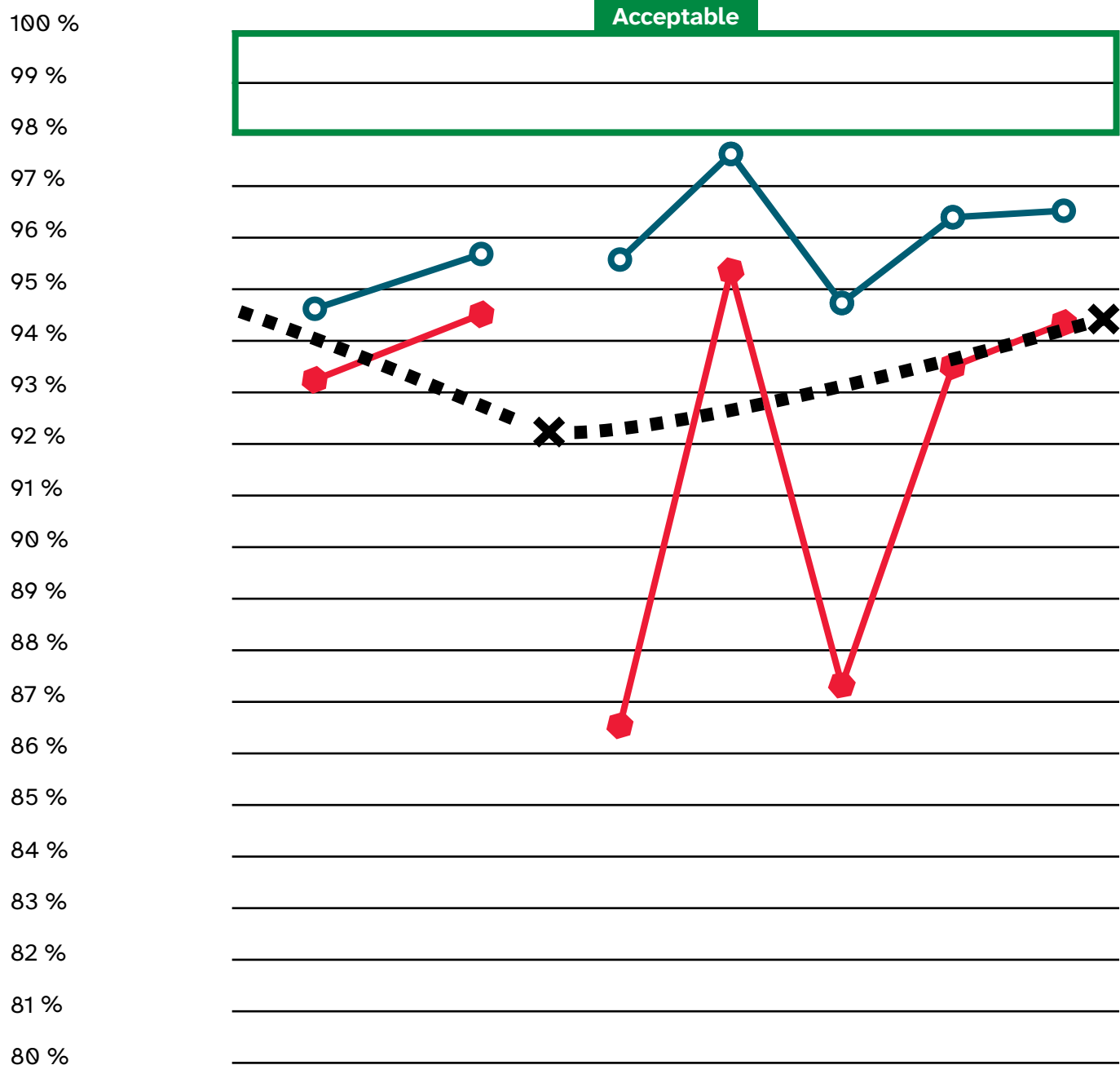
	Nord-occidental				Valencià			
Softcatalà	97,50 %	96,92 %	96,34 %	97,03 %	96,88 %	96,45 %	97,25 %	95,27 %
YouTube	96,94 %	96,02 %	95,48 %	96,37 %	96,47 %	94,83 %	96,68 %	95,00 %
Mitjana del conjunt d'aplicacions	96,74 %				96,30 %			

Figura 16. Puntuació NER per tipus de discurs.



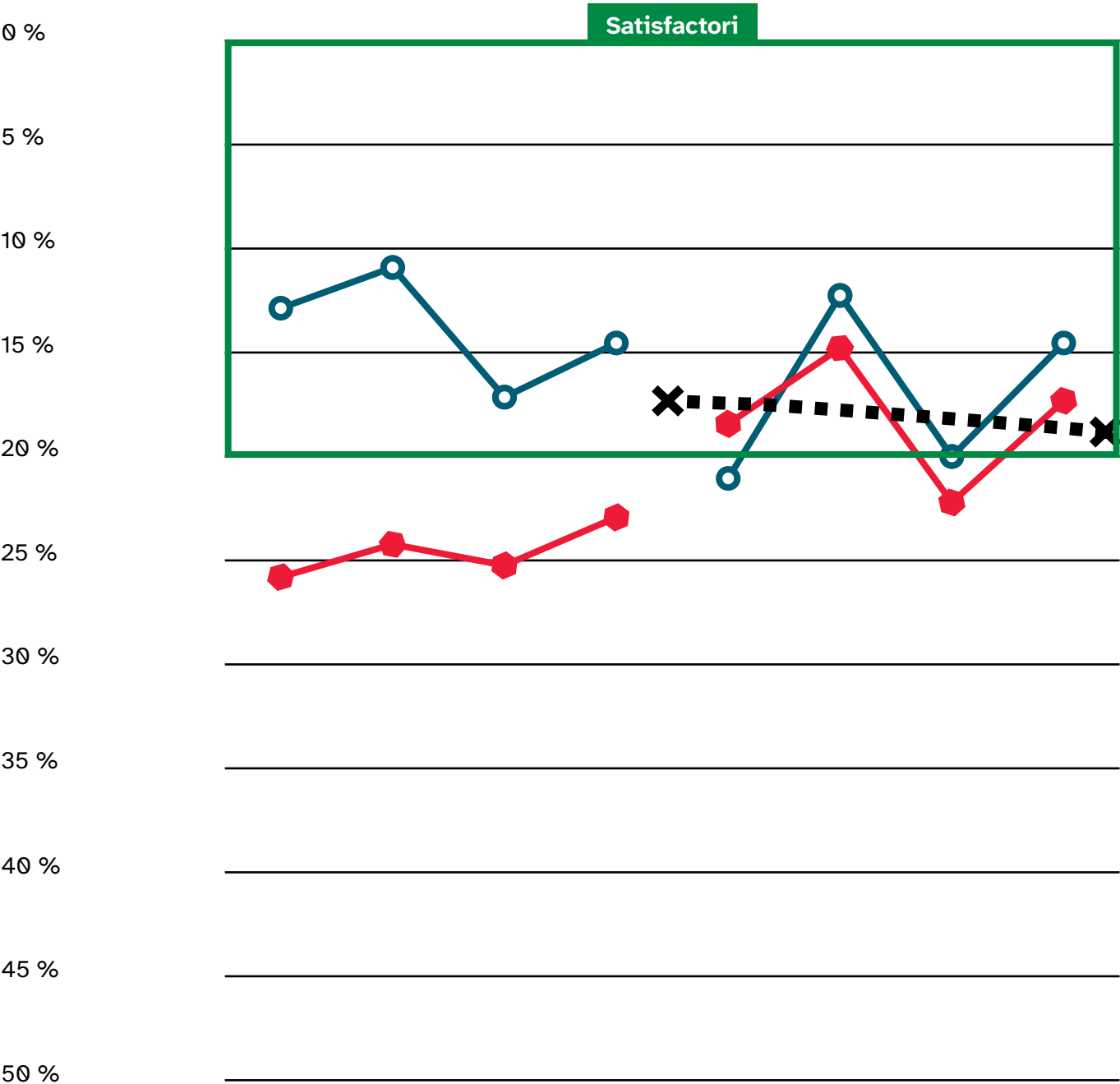
	Central - preparat				Central - espontani			
● Softcatalà	97,09 %	97,45 %	96,39 %	97,94 %	98,23 %	98,39 %	96,82 %	96,03 %
● YouTube	93,13 %	94,08 %	92,59 %	92,68 %	94,81 %	95,75 %	96,52 %	95,77 %
✕ Mitjana del conjunt d'aplicacions	96,54 %				95,91 %			

Font: Elaboració pròpia



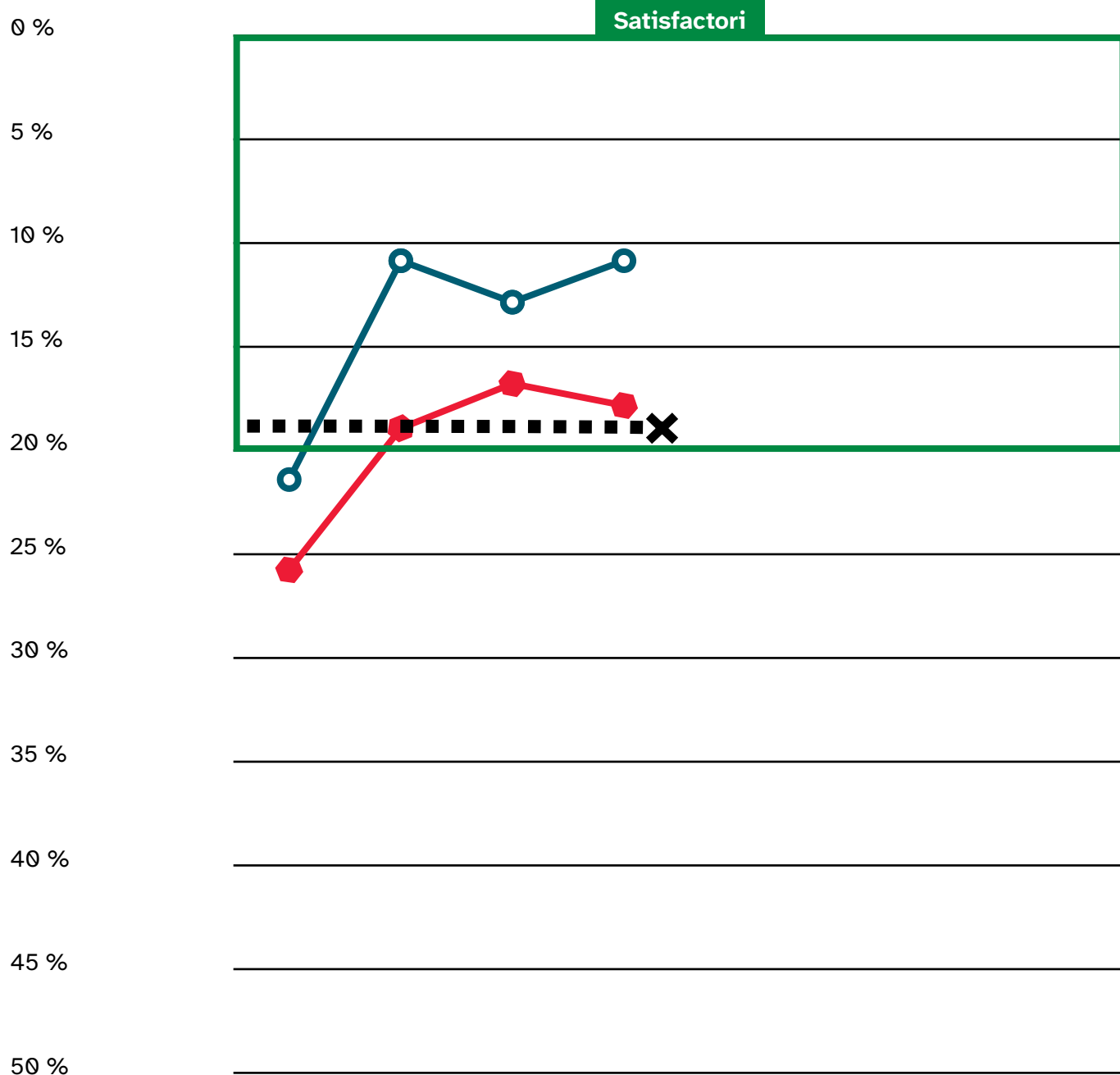
	Central amb interrupcions		Central amb canvi de llengua CA-ES				
● Softcatalà	94,65 %	95,82 %	95,63 %	97,68 %	94,92 %	96,46 %	96,52 %
● YouTube	93,13 %	94,64 %	86,66 %	95,22 %	87,37 %	93,56 %	94,24 %
✕ Mitjana del conjunt d'aplicacions	92,13 %						94,32 %

Figura 17.Puntuació WER per tipus de situació de comunicació no ideal.



	Central - preparat				Mala qualitat de so			
● Softcatalà	12,97 %	10,92 %	17,14 %	14,44 %	21,08 %	12,07 %	20,00 %	14,44 %
● YouTube	25,95 %	24,14 %	25,14 %	22,99 %	18,38 %	14,94 %	22,29 %	17,11 %
✕ Mitjana del conjunt d'aplicacions	17,35 %				18,75 %			

Font: Elaboració pròpia



	Entorn sorollós			
Softcatalà	21,51 %	10,92 %	12,97 %	10,92 %
YouTube	25,81 %	18,97 %	16,76 %	17,82 %
Mitjana del conjunt d'aplicacions				18,85 %

Resultats específics

Si tenim en compte la quantitat i la gravetat dels errors en les situacions de producció ideals, observem que Softcatalà es troba en un punt mitjà de la taula quant a nombre d'errors, però destaca en ser l'eina amb menys de lleus.

Per tipus d'error, la majoria són de reconeixement sense causa concreta —44 %—, seguits, però, dels d'ortografia i convencions —11 %—. En general, es tracta d'errors en l'ús de majúscules i de confusions amb paraules homònimes.

Transcripció manual:

Vam anar a xinesos, vam anar a hindús, vam anar a tot arreu.

Transcripció automàtica:

Vam anar a **Xinesos**, vam anar a **Indús**, vam anar a tot arreu, [...]

Transcripció manual:

Perquè de repent tenc un buit de temps lliure [...]

Transcripció automàtica:

[...] perquè de sobte en tinc un **vuit** de temps lliure [...]

Transcripció manual:

[...] i també destinant una part del meu cervell a pues tot aquell... tot aquell àmbit.

Transcripció automàtica:

[...] i també he destinat una part del meu **servei** a tot aquell àmbit.

En aquest sentit, destaca també el fet que, en les mostres amb barreja de llengua, trobem un nombre similar de casos en què tradueix, en què manté el castellà i en què barreja totes dues

llengües. Per tant, podem observar una tendència general a cometre errors en la distinció de paraules amb una pronúncia propera.

Alhora, també destaca quant al nombre d'errors relacionats amb noms propis —7 % dels seus errors, però 21 % del total dels errors d'aquest tipus—.

Transcripció manual:

Després de Kuala Lumpur vam anar cap a unes illes [...]

Transcripció automàtica:

Després de **Guadalumpur** vam anar cap a unes illes, [...]

Transcripció manual:

Ella va actuar múltiples vegades al Gran Teatre del Liceu de Barcelona.

Transcripció automàtica:

Ella va actuar múltiples vegades al Gran Teatre de **l'Esteu** de Barcelona.

Cal destacar que Softcatalà és l'eina que menys errors de puntuació ha comès, amb només el 4 % dels errors de totes les aplicacions. Així mateix, és l'eina que menys incorreccions ha fet a causa de dubitacions i reformulacions, i també la que menys omissions presenta.

A més, és la segona eina que més reconeixements correctes i correccions ha fet.

Transcripció manual:

Després també m'agrada molt les polaritats, no?

Transcripció automàtica:

Després també **m'agraden** molt les polaritats.

En el cas de YouTube, aquesta eina ocupa la tercera posició en nombre d'errors, globalment i en els tres tipus de gravetat.

Per tipus, la gran majoria d'errors són de reconeixement sense causa aparent —31 %—. En aquest sentit, destaca el fet que és l'única aplicació que transcriu conjunts de caràcters que no formen paraules en català, sinó que més aviat corresponen a una representació fonètica basada en correspondències so-grafia.

Transcripció manual:

I ara mateix estem a dimarts [...]

Transcripció automàtica:

I ara mateix ja estem a **dimars** [...]

Transcripció manual:

[...] “Ja veuràs que et... que et canviarà l'accent”.

Transcripció automàtica:

[...] ja **veoràs** que et canviarà l'accent [...]

En segon lloc, trobem un gran nombre d'errors de puntuació —24 %—, de totes les gravetats.

Transcripció manual:

Vaig ser presentador, director, guionista, productor de un magazín [...]

Transcripció automàtica:

vaig ser **presentador director guionista productor** d'un magazín [...]

Transcripció manual:

[...] i la Història del Món de Contemporani de 1r de Batxillerat.

Transcripció automàtica:

[...] i la història del món de **contemporani**.
Primer de batxillerat.

Cal destacar també que YouTube és l'eina que més errors ha comès en el reconeixement de noms propis —25 % del conjunt de les eines— i xifres —44 %—, la que més addicions o al·lucinacions presenta —24 %— i la segona amb més errors causats per reformulacions del discurs —23 %—. En aquest sentit, els resultats encaixen amb la tendència a representar amb una aproximació fonètica tots els sons que detecta, tot i que no siguin parla o paraules de la llengua.

Transcripció manual:

M'he anat movent molt com de Manacor a Portocolom, [...]

Transcripció automàtica:

[...] anat movent molt com de **menacó porta colom** [...]

Transcripció manual:

Telefona al 650 05 33 98.

Transcripció automàtica:

Telefona al 650**00** 053398.

Transcripció manual:

La veritat és que donen pluges per a aquesta setmana, [...]

Transcripció automàtica:

[...] la veritt és que **no** donen pluses per aquesta setmana, [...]

Softcatalà és l'eina que menys errors de puntuació i a causa de dubitacions i reformulacions ha comès.

Transcripció manual:

Jo vaig marxar a estudiar a Barcelona fa... Farà cosa de... de dos anys.

Transcripció automàtica:

Jo vaig marxar a estudiar a Barcelona **fa la** cosa de de dos anys [...]

Com a aspectes positius, podem destacar que, juntament amb Google Meet, és l'eina que menys ha estandarditzat les varietats dialectals. De la mateixa manera, i tenint també en compte la seva tendència a transcriure aproximacions fonètiques, podem observar que, en aquells casos en què reconeix correctament els termes, acostuma a mantenir els termes en castellà en els textos amb barreja, tot i que en tradueix alguns amb pronúncia molt similar.

Avaluació de les característiques d'accessibilitat

Si tenim en compte altres aspectes rellevants perquè la transcripció sigui accessible, podem destacar que YouTube és l'única aplicació d'aquest grup que identifica els parlants, amb un doble signe de més gran (>>) o un salt de línia, tot i que no sempre. En algunes ocasions, però, el signe apareix just després de la primera o segona paraula de la intervenció del nou parlant. De fet, és l'única eina del conjunt total que, tot i identificar els parlants, comet

errors d'identificació i de barreja i omissió d'intervencions.

Quant a la personalització, YouTube és també l'única eina amb una doble orientació. Com a usuari final que consumeix els subtítols a la plataforma, es poden personalitzar aspectes com la tipografia, el color i la mida de la lletra, o el color i l'opacitat del fons. Com a persona que crea els subtítols, en canvi, no es pot personalitzar cap aspecte, ni de visualització ni de les característiques de la transcripció.

En el cas de Softcatalà, l'eina està orientada a un públic general, que podrien ser tant tècnics d'accessibilitat com persones usuàries. Aquesta eina tampoc no permet personalitzar aspectes de visualització, però aquest fet no es pot considerar negatiu, ja que és l'única aplicació en què els subtítols no es consumeixen dins la plataforma. No obstant això, és també l'única aplicació que permet configurar el nombre màxim de caràcters per línia i de línies per subtítols.

YouTube és l'única aplicació que transcriu conjunts de caràcters que no formen paraules en català, sinó que corresponen a una representació fonètica.

04 • Conclusions

Aquest és el tercer any que fem aquesta recerca —el segon amb aquest nivell d'ambició— i, tot i que les aplicacions, el corpus i les variables analitzades han estat diferents en tots tres casos i no ens permeten oferir dades rigoroses de l'evolució d'aquesta tecnologia, sí que podem parlar d'una evolució en el seu ús per part de la nostra base social i de la societat en general.

Així mateix, durant aquests anys hem vist aparèixer moltes aplicacions noves i incorporar el català a moltes d'altres existents.

D'aquesta manera, els estudis anteriors ens van permetre fer una primera aproximació a aquesta tecnologia i als motors principals de reconeixement de veu en català —així com a algunes de les aplicacions concretes—, per valorar la possibilitat que algun dia fossin eines d'accessibilitat per a persones amb sordesa que es comuniquen oralment. Així, vam poder anar descartant aquelles aplicacions que no ofereixen una qualitat prou alta per plantejar que siguin recursos útils per a persones amb sordesa, i seguir la pista d'aquelles amb potencial d'arribar a ser eines d'accessibilitat.

Recentment, però, la ràpida evolució de les tecnologies del llenguatge i de la intel·ligència

artificial ha accelerat l'ús de la transcripció automàtica entre persones amb sordesa i entitats que s'hi comuniquen. Per tant, la recerca d'enguany no només ha buscat trobar patrons i tendències en abstracte que, sense pretendre ser extrapolables a tots els usos i totes les aplicacions, ens permetin fer-nos una idea de la situació actual d'aquesta tecnologia i de la viabilitat de fer-la servir com a recurs d'accessibilitat, sinó que també ha pres com a objectiu poder dotar els usuaris de criteri tecnològic per a l'elecció i l'ús conscient de cadascuna de les aplicacions concretes.

Com a primera conclusió, l'anàlisi de la qualitat de les transcripcions automàtiques en català d'aquestes set aplicacions mostra que aquesta tecnologia encara no està prou desenvolupada per esdevenir una eina

La transcripció automàtica en català encara no està prou desenvolupada per esdevenir una eina d'accessibilitat en si mateixa que permeti complir les obligacions legals d'accessibilitat a la comunicació.

d'accessibilitat en si mateixa, que permeti complir les obligacions legals d'accessibilitat a la comunicació. En cap cas, per tant, pot substituir altres mesures d'accessibilitat tradicionals, especialment la subtitulació humana.

Tot i que hi ha certes situacions de comunicació en què algunes de les eines obtenen molt bons resultats —discursos preparats en registre estàndard, produïts en entorns de so controlat i enregistrats amb un sistema d'àudio de qualitat—, encara hi ha molta feina per fer en altres aspectes. En aquest sentit, és especialment necessària una millora de l'efectivitat del reconeixement de veu en situacions de comunicació quotidianes i de relacions interpersonals, és a dir, en situacions en què no és factible la contractació de subtitulació humana o l'ús d'altres recursos d'accessibilitat i en què, per tant, les persones amb sordesa que es comuniquen oralment tenen més dificultats d'accés a la informació i comunicació.

Més concretament, cal treballar en la millora del reconeixement de les diverses varietats dialectals del català, tant pel que fa a aspectes fonètics com lèxics i gramaticals, i especialment quant a les varietats balears. D'aquesta manera,

la transcripció automàtica permetrà l'accés a la informació també en situacions de comunicació no formals i sense limitació d'espai dins el territori de parla catalana, alhora que contribuirà a la defensa dels drets lingüístics dels parlants, que no es veuran obligats a canviar de manera de parlar o de llengua vehicular per comunicar-se amb persones amb sordesa.

Així mateix, també cal fer esforços quant a la millora del reconeixement de discursos no ideals, que no han estat preparats i assajats amb antelació i que, per tant, presenten un gran nombre de dubitacions, reformulacions i errors gramaticals, així com d'interrupcions entre parlants. En aquest sentit, cal també destacar una necessitat de millora en la puntuació, que és un dels errors més freqüents d'aquestes eines i una qüestió especialment rellevant en les transcripcions per a persones amb sordesa, que en alguns casos poden no tenir accés a aspectes com l'entonació i dependre exclusivament d'ella per a la comprensió del text.

En aquesta línia, també cal fer èmfasi en la necessitat de treballar en altres aspectes de les transcripcions que són rellevants o necessaris per a persones amb sordesa,

És especialment necessària una millora de l'efectivitat del reconeixement de veu en situacions de comunicació quotidianes i de relacions interpersonals.

abordant la millora de les eines des del punt de vista de l'accessibilitat, i no només d'un punt de vista d'altres àmbits o usos, com s'ha fet tradicionalment. Més concretament, es tracta d'aspectes relacionats amb l'accés a informació metalingüística sobre el discurs i la manera de parlar dels oradors: la correcta identificació dels parlants, la no estandardització dels trets dialectals, el manteniment dels termes en altres llengües —especialment del castellà— o la no supressió de mots crossa. A més, s'hauria de plantejar la possibilitat de detectar i transcriure també elements que no són de la parla —música, aplaudiments, riure i plor, per exemple—. Si més no, tots aquests aspectes haurien de ser configurables, perquè cadascun dels usuaris, amb sordesa o sense, pugui decidir si accedeix o no a aquesta informació a través de la transcripció.

Finalment, tot i que es tracta d'aspectes que afecten el reconeixement en un grau més baix, convé també treballar en el reconeixement del discurs quan aquest es produeix en situacions de comunicació no ideals, millorant els sistemes de reducció de soroll de fons o el reconeixement d'enregistraments de baixa qualitat.

Això no obstant, la transcripció automàtica en català sí que pot ser un recurs de suport per suplir la manca d'accessibilitat per a persones amb sordesa que es comuniquen oralment en certs espais i contextos. En aquest sentit,

convidem les persones amb sordesa, així com totes aquelles persones i entitats que s'hi comuniquen, a analitzar els discursos i les situacions de comunicació que necessiten fer accessibles —i on altres mesures d'accessibilitat no siguin possibles— i fer servir els resultats d'aquest projecte per escollir les eines més adients en cada cas, tenint-ne clars els punts forts i les limitacions.

Com a resum dels resultats, podem destacar que les eines que han donat millors resultats globalment són Speechmatics i AgilText. Totes dues permeten la transcripció en directe i en diferit a través del navegador web, però Speechmatics està més enfocada a perfils professionals de l'àmbit del reconeixement de veu —amb la possibilitat d'incorporar-se a altres eines i amb diverses opcions de configuració d'aspectes tècnics de la transcripció—, mentre que AgilText està enfocada a usuaris i tècnics d'accessibilitat —amb una interfície més amable i amb diverses opcions de configuració d'aspectes de visualització—.

A més, tot i que és una aplicació encara en proves que no assoleix tan bons resultats, cal mencionar el potencial de SCRIBAL, una eina especialitzada en contextos acadèmics que permet la transcripció en directe i diferit a través del navegador web.

La transcripció automàtica en català pot ser un recurs de suport per suplir la manca d'accessibilitat per a persones amb sordesa que es comuniquen oralment en certs espais i contextos.

Les eines que han donat millors resultats globalment són Speechmatics i AgilText.

Fins que aquestes eines que sí que es poden recomanar com a eines de suport assoleixin una qualitat suficient, caldrà seguir algunes pautes per adaptar els discursos quan les fem servir i reduir la probabilitat que cometin errors. En aquest sentit, cal no oblidar totes aquelles mesures que s'apliquen quan es fan servir altres eines d'accessibilitat, inclosa la subtitulació humana, com poden ser parlar a un ritme adequat i amb una dicció clara.

Més específicament, en fer servir transcripció automàtica cal tenir present que el discurs, encara que sigui espontani, ha de ser endreçat i net, procurant acabar totes les frases i evitant mots crossa —*bueno, diguem-ne, i tal, no?, saps?*—. En converses amb més d'un interlocutor, a més, cal respectar especialment els torns de paraula, tot evitant interrupcions i superposicions.

De la mateixa manera, cal procurar que la llengua sigui el més estàndard i normativa possible, reduint sobretot l'ús de paraules inventades, onomatopeïques o derivades —*tardeig, buf, blablabla, superparadisíac, vermutet*—, així com de castellanismes o anglicismes.

També cal tenir present que, si es fan servir noms propis de qualsevol mena o xifres concretes, cal assegurar que s'estan transcrivint correctament i, si s'escau, fer aclariments o correccions. En el cas d'algunes aplicacions, es pot afegir un diccionari amb termes personalitzats prèviament, però si no és possible cal oferir la informació en un suport

escrit alternatiu, en funció del tipus de situació comunicativa —anotació en un paper o al mòbil, fulletó o infografia, presentació de diapositives, etc.—.

Per acabar, cal destacar que, tot i que aquest estudi ha buscat la representativitat i la utilitat en l'elecció tant de les aplicacions com de les variables analitzades, el corpus de mostres ha estat reduït. Així, no hem pogut tenir en compte variables com l'edat del parlant o l'especialització del discurs, de manera que desconexem com aquests aspectes —i molts altres— afecten el reconeixement de veu d'aquestes eines en català. De la mateixa manera, tampoc no hem pogut aprofundir en les diverses subvarietats de les varietats dialectals, en els tipus i intensitats de soroll de fons, o els discursos que barregen altres combinacions de llengües.

A més, la qualitat de les aplicacions de transcripció automàtica com a eines d'ús quotidià per a persones amb sordesa pot avaluar-se a partir d'altres aspectes més enllà del reconeixement de veu, com la velocitat de posada en marxa, la possibilitat de fer-la servir sense connexió o el preu, que no s'han tingut en compte en aquest informe.



OBSERVATORI

DE LA SORDESA, L'ACCESSIBILITAT
I LA COMUNICACIÓ ORAL

Una recerca de:



Famílies i persones amb sordesa

Ha donat suport al projecte:



Generalitat de Catalunya
**Departament
de Política Lingüística**



Generalitat de Catalunya
**Departament
de Cultura**



Generalitat de Catalunya
**Departament de Drets Socials
i Inclusió**



MIXT
Paper | Donant suport
a la silvicultura responsable
FSC® C074301



Famílies i persones amb sordesa

Famílies i persones amb sordesa que parlem
i treballem la inserció laboral, l'educació
inclusiva, l'accessibilitat a la comunicació.

C/Nàpols, núm. 351 baixos
08025 Barcelona
hola@federacioacapps.org