Desafíos en a construzión de corpus paralelos en aragonés: o camino enta ra traduzión automatica neuronal[1]

Alejandro Pardos Calvo*
Universidad de Zaragoza

Resumen El proyecto *Traducción automática neuronal para las lenguas románicas de la península ibérica* (TAN-IBE), financiado por el Ministerio de Ciencia, Innovación y Universidades, tiene como objetivo entrenar sistemas de traducción automática basados en redes neuronales para siete lenguas románicas de la península ibérica, entre ellas el aragonés. Estos sistemas se entrenan mediante corpus paralelos, es decir, recopilaciones masivas de datos alineados que pueden utilizarse con distintos fines. En lenguas como el castellano, el portugués o el catalán —que cuentan con recursos claramente superiores—, la recopilación de esos materiales no plantea grandes dificultades. En cambio, para lenguas como el aragonés, el asturiano o el aranés, esta tarea resulta compleja debido a la escasez de recursos y de herramientas disponibles. En el caso del aragonés, a ello se suman otros obstáculos, como el bajo grado de digitalización, la considerable magnitud de los corpus dialectales o la diversidad de grafías y normas, entre otros. Este artículo analiza las dificultades específicas a las que se enfrentan los investigadores al recopilar o construir corpus paralelos en lengua aragonesa, evalúa los recursos existentes y propone posibles soluciones y estrategias para superar esos obstáculos.

Palabras clave Lengua aragonesa. Corpus paralelos. Traducción automática neuronal. Entrenamiento de sistemas de traducción. *Transfer learning*. Sistemas multilingües.

Abstract The *Neural machine translation for the Romance languages of the Iberian Peninsula* project (TAN-IBE), funded by the Ministry of Science, Innovation and Universities, aims to train neural network-based machine translation systems for seven Romance languages of the Iberian Peninsula, including Aragonese. These systems are trained using parallel corpora, that is, large collections of aligned data used for various purposes. For languages such as Spanish, Portuguese, or Catalan (which have indisputably stronger resources), data collection does not pose major difficulties. However, for languages like Aragonese, Asturian, or Aranese, it remains a demanding task due to the scarcity of available resources and tools. In the specific case of Aragonese, additional challenges arise, including

* a.pardos@unizar.es

1 Comunicazión presentata en a IX Trobada d'Estudios e Rechiras Arredol d'a Luenga Aragonesa e a suya Literatura (Uesca, 3-5 d'otubre de 2024).

limited digitization, the considerable size of dialectal corpora, and the diversity of spellings and norms, among other factors. This paper analyzes the specific difficulties researchers face when collecting or constructing parallel corpora in Aragonese, assesses existing resources, and outlines potential solutions and strategies to overcome those obstacles.

Keywords Aragonese language. Parallel corpora. Training of machine translation systems. Transfer learning. Multilingual systems.

Résumé Le projet *Traduction automatique neuronale pour les langues romanes de la péninsule ibérique* (TAN-IBE), financé par le ministère espagnol des Sciences, de l'Innovation et des Universités, a pour objectif de former des systèmes de traduction automatique basés sur des réseaux neuronaux pour sept langues romanes de la péninsule ibérique, dont l'aragonais. Ces systèmes sont entraînés à l'aide de corpus parallèles, c'est-à-dire de vastes compilations de données alignées pouvant être utilisées à différentes fins. Dans des langues telles que l'espagnol, le portugais ou le catalan —qui disposent de ressources nettement supérieures—, la collecte de ces matériaux ne pose pas de grandes difficultés. En revanche, pour des langues comme l'aragonais, l'asturien ou l'aranais, cette tâche s'avère complexe en raison de la rareté des ressources et des outils disponibles. Dans le cas de l'aragonais, d'autres obstacles s'y ajoutent, tels que le faible degré de numérisation, l'ampleur considérable des corpus dialectaux ou la diversité des orthographes et des normes, entre autres. Cet article analyse les difficultés spécifiques auxquelles sont confrontés les chercheurs lors de la collecte ou de la construction de corpus parallèles en langue aragonaise, évalue les ressources existantes et propose des solutions et des stratégies possibles pour surmonter ces obstacles.

Mots clés Langue aragonaise. Corpus parallèles. Traduction automatique neuronale. Formation des systèmes de traduction. *Transfer learning*. Systèmes multilingues.

A traduzión automatica neuronal (TAN), alazetata en retes neuronals fundos, ha surtito en os zaguers años como una teunolochía abanzata ta traduzir entre diferens idiomas con buena prezisión e fluidez. En contimparanza con atros sistemas de traduzión, como aquers que s'alazetan en regles u en frases, os modelos de TAN emplegan un aprendizache fundo ta analizar grans cantidaz de testos, asimilando patrons complexos en as estruturas gramaticals e semanticas d'as luengas. Ista enfiladura premite que os sistemas de traduzión automatica replequen matizes, senditos e contestos que se perdeban en atros metodos anteriors, chenerando asinas traduzions más naturals e adecuatas ta ra situgazión (Genabith, 2020).

A más gran parti d'o desarrollo e entrenamiento de modelos de TAN s'ha zentrato en luengas mayoritarias como l'anglés, o castellano, o franzés u o chino, en do sobrexen os recursos lingüisticos e os corpus paralelos de calidá e no bi ha problemas de finanziazión. Manimenos, ista enfiladura dixa de banda muitas atras luengas minoritarias que, á ormino, no poseden prous datos de calidá ta entrenar ditos modelos. Por tanto, os sistemas de traduzión disponibles no consiguen replecar as carauteristicas distintibas d'istas luengas, perpetuando ra suya imbisibilidá dichital e contrebuyindo á ra suya marguinazión en o contesto teunolochico (Oliver, 2020a). Ta ras luengas minorizatas como l'aragonés, a TAN s'amuestra alazetal ta ra suya preserbazión e ra suya promozión.

L'aragonés, con una comunidá de fablans redueita e una presenzia encara limitata en os meyos de comunicazión, asinas como en os entornos dichitals, se troba

en una situgazión de bulnerabilidá soziolingüistica que incrementa ro risque de desplazamiento e, á ro largo, de desaparizión (Eito *et alii*, 2025). Fren á isto, o desarrollo de modelos de prozesamiento d'o lenguache natural entrenatos de propio ta l'aragonés ubre una nueba linia de treballo con potenzial impauto en diferens libels: (i) puede millorar a disponibilidá de ferramientas de traduzión e de produzión testual, (ii) enfortir a bisibilidá e ro emplego d'a luenga en espazios dichitals de comunicazión cutiana (como en retes sozials e blogs) e (iii) fazilitar a chenerazión de recursos didauticos e materials ta l'amostranza e l'aprendizache. Antiparti, a esistenzia d'uns alazez lingüistico-teunolochicos en aragonés puede contrebuyir á ra suya incorporazión en ambitos instituzionals e almenistratibos, con efeutos positibos sobre a lechitimidá perzibita d'a luenga e a suya funzionalidá en rechistros formals (Baxter, 2017; Busquets, 2020).

Ta que ra traduzión automatica neuronal pueda ufrir toz os suyos benefizios á ras luengas minoritarias, ye alazetal poder contar con recursos lingüisticos adecuatos, estando ros corpus paralelos uno d'os elementos prenzipals en iste prozeso. Por *corpus paralelos* entendemos un conchunto de testos aliniatos en dos u más luengas do cada segmento (troz) en una luenga tiene una traduzión equibalén en l'atra luenga (Gómez, 2005). Istos segmentos furnen datos contimparables que aduyan á ros modelos á replecar as relazions lingüisticas entre as diferens luengas. Ditos modelos aprenden, de traza organizata, cómo se traduzen as parolas e as esprisions en determinatos contestos e deteutan patrons de correspondenzia e equibalenzia entre as luengas en custión. Esenzialmén, un corpus paralelo autúa como una base de datos bilingüe (u multilingüe) en que cada troz u segmento tiene ra suya traduzión correspondién en atra luenga, o que fazilita ra obtenzión de regles de traduzión ta ros modelos de TAN (Castillo, 2011). En o que pertoca á ra tematica, istos gosan tener una gran bariedá de cheneros e estilos (literatura, notizias, decumentazión teunica, ezetera), furnindo asinas tot un bentallo de rechistros e contestos d'emplego, o que constituye una representazión d'a bariabilidá lingüistica pro ampla.

Como s'esplanicaba en alto, á penar d'a importanzia d'os corpus paralelos en o desarrollo d'a traduzión automatica neuronal, no pas totas as luengas cuentan con a mesma cantidá de recursos. Mientres que en as luengas mayoritarias istos corpus son amplos e cubren gran bariedá de contestos e cheneros, en luengas minoritarias como l'aragonés a reyalidá ye bien diferén. A manca de corpus paralelos de calidá representa un barrache en a creyazión de modelos de TAN que sigan prezisos, dato que limita ras posibilidaz de traduzir ista luenga fidelmén e funzional en diferens contestos. En o presén articlo s'analizará ra situgazión de l'aragonés en o tocán á ra disponibilidá e ra calidá d'os corpus paralelos, rancando d'a esperenzia replegata en o proyeuto TAN-IBE (*Traducción automática neuronal para las lenguas románicas de la península ibérica*) (Oliver, 2020b).

Dende abiento de 2022 dica mayo de 2023, en o proyeuto TAN-IBE, coordinato por a Universitat Oberta de Catalunya e finanziato por o Ministerio de Ciencia, Innovación y Universidades, se treballó en a replega, l'analís, a edizión e a porga

de corpus paralelos en aragonés (asinas como en as atras luengas romanicas minoritarias que fan parti d'o proyeuto), prenendo ro castellano como luenga de referenzia. Una güellada inizial dixaba alufrar que l'aragonés partiba d'una clara desbentalla fren á ra resta d'as luengas romanicas d'a peninsula iberica. Seguntes a web Opus Corpus, referenzia clau en a replega de datos lingüisticos, l'aragonés se posiziona como a luenga romanica peninsular con menos segmentos disponibles, nomás por denzima d'o mirandés, que en beras fa parti d'o diasistema asturleonés, pero dispone d'un repositorio propio. Con alto u baxo sisanta mil segmentos disponibles, se troba muito luen de luengas minorizatas como l'occitano, que cuenta con dos millons de segmentos, u l'asturiano, que, á penar d'a suya situgazión, tamién delicata, dispone de más de treze millons. Goyan de millor salú encara atras luengas, como ro catalán, o gallego u, profes, o portugués.

Portugués	482 900 000
Catalán	163 700 000
Gallego	67 900 000
Asturiano	13 300 000
Aranés (occitano)	2 000 000
Aragonés	60 700
Mirandés	9 200

Tabla 1. Lumero de segmentos d'os corpus paralelos entre o castellano e ra resta d'as luengas disponibles en a colezión d'Opus Corpus. (Fuen: Elaborazión propia)

O prozeso de replega de corpus paralelos s'enzetó de traza organizata e con criterios esclateros ta guarenziar una obtenzión de datos efeutiba. Asinas, s'establioron tres linias de treballo estratechicas: se prebó de coordinar esfuerzos con as diferens instituzions poseyedoras de material bilingüe; s'optó por a replega indibidual e por a creyazión de nuebos corpus paralelos por parti d'o equipo rechirador; se prozedió á ra clasificazión d'os datos en dembas tematicas que respondesen á ros diferens contestos d'emplego.

En un primer inte se contautó con as instituzions e as entidaz prenzipals de l'ambito de l'aragonés, encluyendo-bi editorials, asoziazions culturals, prensa, meyos de comunicazión e creyadors de contenito, entre atros, con l'obchetibo d'identificar posibles fuens de corpus paralelos aragonés-castellano. As respuestas estioron bariatas: bellas entidaz colabororon de traza faborable, atras refusoron a suya partezipazion u determinoron de no responder. En iste causo, cal acobaltar o compromís que adotó ra Direzión Cheneral de Politica Lingüistica, que, de camín, furnió

1 Como no bi'n eba de corpus en aranés, emos establito ras zifras de l'occitano.

una gran bariedá de materials, tanto bilingües como monolingües, que constituyoron a base tematica más dibersa de tot o prozeso de replega. Amás d'o contenito textual, a instituzión fazió partache d'atros recursos de gran balura, como ro contenito de l'*Aragonario*, dizionario web castellano-aragonés, e barios mils de lexemas que eban paratos ta puyar ta Wikidata, enamplando asinas os recursos disponibles ta desarrollar a TAN en ista luenga.

De bez, o presonal rechirador prozedió á ra replega direuta de corpus paralelos obtenitos d'atras fuens. S'adibioron recursos como nobelas, libros de diferens tematicas u materials didauticos dichitalizatos, pachinas web bilingües, contenitos de blogs e retes sozials, articlos de prensa, rechiras academicas, ezetera. Como nobedá cal acobaltar a incorporazión de toz os articlos de Wikipedia disponibles en aragonés, que constituyen buena parti d'os recursos dichitals, encara que en bel causo presentan inconsistenzias en cuanto á calidá u equibalenzia, dato que muitos son estatos cheneratos automaticamén dende ro castellano u dende atras luengas. Ta compensar istas limitazions e refortir as bases de datos s'optó tamién por a traduzión de nuebos corpus disponibles en castellano, como ro *Flores-200*, que furnió mils de nuebos segmentos. A rebisión d'as alportazions por parti d'os espertos en cada una d'as luengas ufre una mayor prezisión lingüistica e millora a calidá cheneral d'os datos disponibles.

Español	402 430
Portugués	429 730
Catalán	133 214
Gallego	39 627
Asturiano	11 734
Aranés (occitano)	14 584
Aragonés	10 552
Mirandés	2 000

Tabla 2. Lumero d'articlos de Wikipedia disponibles en cada luenga. (Fuen: Elaborazión propia)

Os datos replegatos estioron triatos inizialmén en dos grans categorías: corpus bilingües e corpus monolingües. Como s'ha endicato antis, os corpus bilingües son alazetals ta ro proyeuto, dato que premiten o entrenamiento direuto d'os sistemas de traduzión automatica furnindo exemplos esclateros d'equibalenzias frase á frase entre aragonés e castellano. Os corpus monolingües, encara que no contrebuyen direutamén á ra creyazión de traduzions, constituyen un recurso de gran balura ta l'analís de patrons de gramatica, lesico e estilo propios de l'aragonés, o que complementa ra capazidá d'o sistema ta capir e chenerar contenitos más naturals en ista luenga (Sennrich, Haddow e Birch, 2015). Por ixo se determinó de replegar corpus monolingües como complemento esenzial ta totas as luengas d'o proyeuto. Ta

masimizar l'aplicabilidá d'o corpus replegato se lebó a cabo una clasificazión tematica que premitió guarenziar en buena mida ra dibersidá contestual. Asinas, s'obtenioron corpus en as dembas de dreito, almenistrazión, literatura, istoria e cultura, educazión, naturaleza e meyoambién.

Durando ro prozeso surtioron prous dificultaz. Muitas d'ellas se poderban atrebuyir á ro contesto cheneral d'as luengas minoritarias; manimenos, s'obserboron bels barraches que afeutan de traza espezifica á l'aragonés. Contino se presentan os prenzipals problemas identificatos en ista fase d'o proyeuto.

1. **Carenzia de testos bilingües**. A propia inesistenzia de testos bilingües aragonés-castellano representa o mayor barrache. D'o total de testos replegatos, os corpus paralelos representan nomás o 47,62 %, mientres que ros corpus monolingües representan o 52,38 %. Sin d'una base zereña de corpus paralelos con correspondenzias esclateras, os modelos de TAN enfrontinan grans problemas.
2. **Estandarizazión lingüistica**. Una proporzión sinificatiba d'o total d'os testos disponibles corresponde á ras diferens barians dialeutals de l'aragonés. D'a mesma traza, a manca d'un modelo lingüistico unificato, emplegato de manera consensuata, dificulta ra creyazión de corpus omochenios. Isto condiziona tanto l'aliniazión de testos entre as dos luengas como ra consistenzia d'as traduzions obtenitas.
3. **Diferens combenzions ortograficas**. A inconsistenzia en a escritura de l'aragonés, resultata d'a coesistenzia de barias combenzions ortograficas, dificulta nuebamén a creyazión de corpus uniformes. Ista bariabilidá chenera confusión en os sistemas d'aliniazión e reduze ra fiabilidá d'os modelos de TAN, que requieren d'una ortografía coderén ta masimizar a prezisión en o prozesamiento de datos.
4. **Manca de dichitalizazión**. Muitos d'os testos escritos en aragonés encara no son dichitalizatos o se troban en formatos difizilmén prozesables, mientres que ra replega automatizata e ro prozesamiento de datos arrienda son aspeutos esenzials ta entrenar modelos de TAN de traza efizién.
5. **Limitazión de recursos web**. Encara que l'aragonés aiga incrementato ra suya presenzia en o rete durando ras zagueras añadas, ista contina estando pro radita. Ista reyalidá limita l'azeso á potenzials fuens de gran calidá en do se chenera un influmen de contenitos autualizatos e adautatos á ro contesto dichital, representans, asinas mesmo, d'una gran bariedá de rechistros.
6. **Calidá d'as traduzions**. As traduzions disponibles en aragonés barían muito en cuanto á calidá. S'obserba una gran diferenzia, por exemplo, entre as traduzions elaboratas durando ras primeras decadas d'a renaxedura d'a luenga e ras más autuals, ya que en ixa etapa ras tendenzias d'emplego

encara no yeran definitas; antiparti, a luenga se trobaba en un inte d'adautazión á ras nuebas nezesidaz. Con tot e con ixo, bi ha muitas traduzions, fueras d'o rechistro emplegato, que no poderban estar consideratas material confitable, dato que son traduzions no profesionals, reyalizatas por presonas afizionatas, e se trata d'un proyeuto que requiere prezisión.

7. **Limitazión en a bariedá d'os rechistros e dembas tematicas**. L'aragonés ha desarrollato amplamén a suya literatura en as zagueras decadas. Manimenos, bi ha una gran manca en a produzión e a traduzión de testos teunicos, zientificos e legals, o que reduze l'aplicabilidá d'os modelos de TAN en contestos formals e espezializatos e dixa ra luenga en una posizión d'inferioridá en contimparanza con atras luengas más consolidatas.
8. **Sobrefaxo de recursos lesicograficos**. Esisten lumerosos recursos lesicograficos, como dizionarios e replegas lesicas, que fan buena onra ta establir equibalenzias prezisas de parolas e ta enamplar o conoximiento d'o bocabulario disponible en dita luenga. Manimenos, istos recursos, fren á atros corpus d'uso prautico, restrinchen o entrenamiento de modelos en contestos reyals d'emplego d'a luenga.
9. **Restrizions d'azeso e dreitos d'autor**. Muitos materials son amparatos por dreitos d'autor, o que limita ra suya disponibilidá e ra suya incorporazión como corpus monolingües u paralelos. Á isto cal adibir a disparidá en a boluntá de colaborazión por parti d'os diferens autors implicatos, que incrementó ras dificultaz de replega.

Dimpués d'analizar con ficazio as dificultaz asoziatas á ra creyazión de corpus paralelos ta l'aragonés, cal planteyar azions que premitan sobrepuyar ista situgazión. Bi ha un dople obchetibo, que pasa por (i) enamplar a disponibilidá de corpus bilingües (e monolingües) de calidá e (ii) enfortir as ferramientas teunolochicas que premitan o desarrollo de modelos de TAN efiziens.

1. **Replega de nuebos testos**. Como ye lochico, o medro d'o corpus aragonés pasarba en primer puesto por a identificazión e a replega de nuebos testos. Isto encluye a esplorazión de nuebas fuens literarias, dichitals e teunicas. A incorporazión de nuebos materials enampla ro rango de rechistros e estilos lingüisticos e premite entrenar modelos más zereños e adautables. Amás, a recopilazión sistematica contrebuye á ra creyazión d'una base de datos más representatiba d'os diferens emplegos de l'aragonés, que comprenda tanto ros contestos más informals como ros academicos u zientificos. Dato que muitas obras no son dichitalizatas, se suchiere priorizar a suya combersión enta formatos azesibles por meyo de teunicas de dichitalizazión como l'OCR (reconoximiento optico de caracteres).
2. **Colaborazions con autors e traductors locals**. O esito de cualsiquier iniziatiba de replega de corpus pende en gran mida d'a colaborazión autiba

con os prenzipals autors d'o ecosistema lingüistico. En iste sendito, ta obtener materials ineditos u poco azesibles estarba alazetal establir alianzas con as editorials locals, asinas como con os autors e traductors más autibos. Istas colaborazions enamplan o bolumen de datos disponibles e guarenzian un menimo de calidá d'os testos replegatos, ya que os profesionals d'a luenga gosan estar millor paratos ta produzir traduzions coderens e fidels. Amás, a implicazión d'istos autors faborexe una mayor conzenzia e un mayor compromís en o desarrollo de ferramientas ta l'aragonés, creyando una sinerchia positiba entre ra lingüistica e ra industria cultural.

3. **Plataformas de *crowdsourcing* ta ra replega de testos**. O *crowdsourcing* constituye una soluzión efeutiba e economica ta xamplar os corpus en o causo d'as luengas minoritarias (Capdevila, 2012). Á trabiés de diferens plataformas dichitals se puede embrecar á ra poblazión ta que furna traduzions, transcrizions u rebisions lingüisticas. S'espleita asinas a colaborazión comunitaria, pro útil ta luengas como l'aragonés, que tiene recursos raditos, pero cuenta con una comunidá de fablans autiba que puede contrebuyir en gran mida. Manimenos, cal implementar mecanismos de balidazión ta guarenziar una calidá e una coderenzia menimas en os materials furnitos. Una estratechia complementaria pasarba por establir combenios de colaborazión con instituzions d'educazión superior, como a Unibersidá de Zaragoza, que imparte un Diploma d'Espezializazión en Filolochía Aragonesa. Ista mena de colaborazions premite canaliar a esperenzia academica e guarenzia un fluxo contino de materials de calidá menima, elaboratos baxo superbisión dozén e con criterios filolochicos zereños, mantenendo a sostenibilidá d'os esfuerzos de replega á ro largo.

4. **Traduzión de testos teunicos, zientificos e legals**. A traduzión d'ista mena de materials ye alazetal ta enriquir os corpus paralelos de l'aragonés, ya que istos rechistros espezializatos no gosan estar guaire representatos en o contesto d'as luengas minoritarias. Por exemplo, en l'ambito almenistratibo, a Comunidá Autonoma d'Aragón chenera un buen influmen de decumentazión legal e teunica como decretos, ordenanzas, resoluzions y autas que poderban traduzir-sen enta l'aragonés. Tamién, en l'ambito pribato, poderban redautar-se contratos estándar u reglamentos que reflexen os diferens conzeutos churidicos en amas luengas. Suzede ro mesmo con atros seutors, como l'agricultura, a chestión ambiental u ra incheniería, dembas relebans en a economía aragonesa, que produzen manuals, guías e estudios teunicos que poderban estar traduzitos á l'aragonés. Finalmén, se poderban encluyir testos binclatos á ro patrimonio cultural e l'autibidá turistica. Materials como guías turisticas, catalogos, descrizions de rutas u folletos intormatibos pueden alportar terminolochía espezializata de bez que posizionan l'aragonés como luenga de cultura e conoximiento en un contesto d'interpretazión patrimonial propia (Pardos e Sanagustín, 2022).

5. **Transcrizión e traduzión de rechistros orals**. L'aragonés tiene una tradizión oral pro ampla que, bien emplegata, poderba constituyir un recurso de gran balura ta ros corpus lingüisticos. A transcrizión e traduzión de rechistros orals como entrebistas, relatos, canzions u atras composizions furnirba un buen exemplo d'emplego d'a luenga en contestos cutianos.

6. **Desarrollo de correutors ortograficos, lematizadors e analizadors espezificos ta l'aragonés**. A creyazión de ferramientas lingüisticas espezificas ye alazetal ta prozesar e normalizar os testos replegatos. Isto premite estandarizar a ortografía, resolber as inconsistenzias identificatas e estruturar os testos en un formato adecuato ta ra suya incorporazión como corpus paralelos. Estarba espezialmén útil, ya que l'aragonés cuenta con diferens combenzions ortograficas e barians dialeutals que, en a prautica, pueden comprometer a unificazión d'os datos lingüisticos.

7. **Aplicazión de teunicas d'aprendizache que profiten datos de luengas romanicas zercanas**. L'aragonés, como luenga romanica, poseye carauteristicas estruturals e lesicas comuns con atras luengas d'a suya familia, como ro catalán. Ista prosimidá lingüistica puede estar espleitata meyán teunicas de transferenzia d'aprendizache (*transfer learning*) e modelato multilingüe (Nguyen e Chiang, 2017). D'ista traza, se pueden entrenar modelos multilingües que encluyan grans corpus de catalán u gallego ta millorar o rendimiento d'a traduzión automatica en aragonés.

O desarrollo de corpus paralelos ta l'aragonés ye una fayena estratechica de gran complexidá, que traszende l'ambito teunico e se fica en o nuclio d'os esfuerzos por rebitalizar ista luenga minoritaria. En o presén articlo s'han identificato tanto ras limitazions estruturals que enfrontina iste prozeso como ras oportunidaz de millora que poderban abordar ditas carenzias. Entre ros prenzipals barraches se puede acobaltar a escasez de testos bilingües e ra manca d'estandarizazión lingüistica, que dificultan l'aliniazión prezisa de materials e comprometen a consistenzia d'as traduzions automaticas. Istos desafíos son, en muitos causos, inerens á ra reyalidá d'as luengas minoritarias, pero se beyen agrabatos por custions que afeutan particularmén á l'aragonés, como ra coesistenzia de diferens combenzions ortograficas u ra presenzia radita en o entorno dichital. Á penar de tot ixo, os abanzes reziens en ferramientas dichitals e metodolochías de prozesamiento de luengache ubren nuebos caminos enta ra recuperazión e ra superazión d'istas dificultaz.

As estratechias proposatas miran d'empentar nuebos paradigmas en as trazas de replegar, prozesar e emplegar datos lingüisticos en aragonés. As iniziatibas más importans pasarban por a colaborazión con instituzions culturals e academicas, a creyazión de plataformas de *crowdsourcing* e l'aplicazión de teunicas d'aprendizache transferito dende atras luengas romanicas zercanas. Amás, a dibersificazión tematica e ra incorporazión de rechistros espezializatos como ros testos zientificos, legals u educatibos resultan alazetals ta guarenziar que ros modelos puedan responder á

ras nezesidaz prauticas e cutianas d'os fablans e ras demandas de ziertos seutors, de bez que premiten enfortir o prestichio d'a luenga en contestos contemporanios.

L'aragonés difizilmén podrá abanzar en rebitalizazión e normalizazión sin corpus paralelos. En un entorno cada begata más dichitalizato, constituyen una ferramienta que premite treballar con rigor e continidá. O esito penderá, en gran mida, d'a coordinazión entre autors locals, instituzions e espertos en prozesamiento d'o luengache natural, como se ye fendo dende ro proyeuto TAN-IBE. Empentar ista mena de propuestas premitirá consolidar l'aragonés como una luenga de raso funzional en o entorno dichital e ficará ros alazez ta futuras autuazions destinatas á ro suyo desarrollo en o sieglo XXI.

Bibliografía

Baxter, Robert Neal (2017), «The importance of interpreter training for minority languages: an analytical overview of the co-official languages in Spain», *Quaderns*, 24, pp. 151-177.

Busquets, Blanca (2020), «Per què és important la traducció de llengües minoritàries?», en *Arts i Humanitats*, blog de l'Universitat Oberta de Catalunya <https://n9.cl/kq2fh> [consulta: 15/5/2020].

Capdevila Fernández, Cristian (2012), «Crowdsourcing i traducció / localització: una amenaça o una oportunitat?», *Revista Tradumàtica*, 10, pp. 237-243.

Castillo Rodríguez, Cristina (2011), «La alineación de un corpus paralelo multilingüe: propuesta de fases para la didáctica de traducción especializada inversa», *Cadernos de Tradução*, 1 (27), pp. 117-142 <https://dialnet.unirioja.es/descarga/articulo/4925266.pdf>.

Eito Mateo, Antonio, José Ángel Iranzo Sanz, Chaime Marcuello Servós e Alejandro Pardos Calvo (2025), *Charrando aragonés: la lengua aragonesa en su zona de uso predominante*, Zaragoza, PUZ.

Genabith, Josef van (2020), «Neural Machine Translation», en Jörg Porsiel (ed.), *Maschinelle Übersetzung für Übersetzungsprofis: Sammelband*, Berlín, BDÜ.

Gómez Guinovart, Xavier (2005), «Procesamiento y aplicaciones de los corpus paralelos», *Novática*, 175, pp. 50-54.

Nguyen, Toan Q., e David Chiang (2017), «Transfer learning across low-resource, related languages for neural machine translation», en Greg Kondrak e Taro Watanabe (eds.), *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Taipéi, Asian Federation of Natural Language Processing, pp. 296-301.

Oliver, Antoni (2020a), «Human translation and machine translation: specificities, uses, advantages and disadvantages», *Linguapax Review*, 8, pp. 111-131 <https://n9.cl/d3odn>.

— (2020b), «Traducción automática para las lenguas románicas de la península ibérica», *Studia Romanica et Anglica Zagrabiensia*, 65, pp. 367-375.

Pardos Calvo, Alejandro, y Victoria Sanagustín Fons (2022), «Turismo cultural y lengua aragonesa: un estudio de caso en el municipio de San Juan de Plan (valle de Chistau, Huesca, Aragón)», *ROTUR*, 16 (2), pp. 186-206 <https://revistas.udc.es/index.php/rotur/article/view/rotur.2022.16.2.8995/g8995_pdf>.

Sennrich, Rico, Barry Haddow e Alexandra Birch (2015), «Improving neural machine translation models with monolingual data», en Katrin Erk e Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlín, Association for Computational Linguistics, pp. 86-96.