Un corpus pal asturianu.

Les tecnoloxíes llingüístiques na consolidación de les llingües minorizaes

Roser Saurí Colomer
Computer Science Department
Brandeis University
roser@cs.brandeis.edu

> "*The old quip attributed to Uriel Weinreich, that a language is a dialect with an army and a navy, is being replaced in these progressive days: a language is a dialect with a dictionary, grammar, parser and a multi-million-word corpus of texts-and they'd better all be computer tractable.*"
>
> Nicholas Ostler
> Foundation for Endangered Languages[1]

1 Introducción

En cuantes que tresmisores d'una visión del mundu y de la historia, el poder de les llingües como aglutinantes sociales y espresión d'una identidá particular ye tan fuerte que les vuelve bien de veces nun problema pa los gobiernos d'imperios o entidaes polítiques multillingües, que de cutio tomen midíes p'amenorgar la diversidá: por exemplu, arrequexando la comunidá llingüística y culturalmente dixebrada nun apartaz de la realidá que se construi como país; cuando creando una tendencia a la marxinación social de los individuos y los pequeños grupos más marcadamente disidentes.

En primer casu, la llingua minorizada tuvo la posibilidá de sobrevivir hasta anguaño, apartada a un mundu fechu esplícitamente de menos pola parte dominante y que precisamente por esto pudo caltener la so estructura social. Fai'l casu de munches de les llingües indíxenes del continente americanu. En segundu casu, irónicamente la llingua pasa de ser un problema pal poder dominante a ser un problema pa los sos falantes: la contina penalización que reciben por usar la so llingua, dende los primeros años d'escolarización hasta los ámbitos profesionales y sociales de la vida adulta, fai qu'ésta dexe de ser el so mediu de comunicación natural y se vuelva una mena de desgracia de la qu'avergoñase, como ún d'esos alifaces familiares que-y toquen a ún ensin lo pidir. A diferencia del casu anterior, esta situación ye davezu propiciada cuando hai un nivel de contactu mayor ente les dos cultures.

La vitalidá d'una llingua ponse entós en peligru cuando se fai una torga pa los sos falantes. Esta ye, llamentablemente, la situación na que s'afayen anguaño munches de les llingües que se falen en mundu, incluyíes les qu'hasta ahora sobrevivieren escaecíes en

[1] Por cuenta de la reseña del *Workshop on Language Resources for European Minority Languages*, que tuvo llugar na *International Conference on Language Resources and Evaluation*, LREC 1998.

sociedaes consideraes arcaiques y ensin capacidá de futuru. Ello débese al contestu globalizador nel que nos alcontramos, qu'amiesta un nivel adicional de diglosia onde yá esistía ún previu porque, ente munches otres coses, impón un procesu d'estandarización cultural que pasa tamién pela creación d'una norma d'usu llingüísticu a favor de les llingües consideraes de más prestixu; esencialmente, les poques con capacidá de trescendencia a los sistemes d'información y con presencia nel desenvolvimientu tecnolóxicu occidental.

Sicasí, la mesma presea cola que s'establez esti procesu unificador ye la d'abrir la posibilidá de garantizar la supervivencia de les llingües minorizaes. La popularización de les tecnoloxíes de la información y comunicación telemática abriéron-y camín al marcu actual de la llamada sociedá de la información. De la mesma, l' accesu a los medios de comunicación de mases, per un sitiu, y a les tecnoloxíes pal procesamientu automáticu del llinguaxe humanu, per otru, permitirán en bona midida la normalización social de les llingües minorizaes, en cuantes que les autorizarán como llingües afayadices pa situaciones comunicatives prestixaes. El desenrueldu de ferramientes y recursos pal procesamientu del llinguaxe ye, poro, un arma pa la supervivencia y la relevancia de les comunidaes llingüístiques minorizaes.

El presente artículu céntrase na rellación potencial de la llingua asturiana col segundu d'estos preseos: les tecnoloxíes llingüístiques. El puntu de partida que m'afala a escribilu ye l'enfotu en que l'asturianu, a diferencia d'abondes llingües del mundo con un númberu reducíu de falantes y en situación de subordinación respecto a otra, entá ta a tiempu d'entrar tamién nel dominiu de les tecnoloxíes llingüístiques. Ello débese, d'uno, a la so contestualización dientro del marcu occidental, que-y garantiza'l nivel d'accesu tecnolóxicu necesario y la disponibilidá mínima de recursos, tanto materiales como humanos; y d'otro, a la esistencia d'un contestu actual mínimamente favorable a la normalización de la llingua.

Otres llingües minorizaes, xeográfica y políticamente averaes al asturianu, ya entraren nesti terrén: el gallegu, l'euskera y el catalán. Les condiciones de partida d'éstes foron relativamente más favoratibles que les de la llingua asturiana, por mor del so estatus de co-oficialidá dientro del marcu políticu nel que s'alcuentra la mayor parte del so territoriu y fala. Sicasí, la esperiencia en toes elles apurre un bon puntu de referencia pa la entrada del asturianu a les tecnoloxíes del llinguaxe. Siguiendo estos trabayos previos a lo mesmo que la esperiencia algamada n'otres comunidaes llingüístiques, vemos que l'inquiz pa la entrada d'una llingua a estes tecnoloxíes ye la construcción d'un corpus llingüísticu. Nesti artículo plantego, entós, la posibilidá d'un proyecto talu pal asturianu.

Naturalmente, la entrada d'una llingua nes tecnoloxíes llingüístiques nun ye la bulda papal que-y asegura la salvación ensin más. L'induldable interés qu'esti campu tien pa les llingües minorizaes ye resultanza de les posibilidaes de desendolcu tanto d'investigación como de preseos concretos de base llingüística que permite: dende diccionarios nel so tradicional formatu de llibru pa un públicu xeneral hasta programes de traducción automática. Poro, la estaya 2 introduz al llector en campu de la Llingüística de Corpus y el so oxetu d'estudiu, y analiza les posibilidaes qu'ofrez dende una perspectiva puramente científica. La estaya 3 fai repás d'esperiencies previes en proyectos de corpus que pueden servir de referencia: d'uno, les pioneres desendolcaes pal inglés; d'otro, les entamaes nel ámbitu xeográficu y políticu del asturianu. Nos dos casos trátase d'estayes bonamente prescindibles pa los llectores que yá tean familiarizaos col campu. Sicasí,

abultábame que l'enclín divulgador del mio artículu obligaba a entrar mínimamente nestos aspectos básicos. A lo cabero, voi afondar na rellación ente llingües minoritaries y tecnoloxía llingüística que s'apuntó enantes (estaya 4.1), y acabaré analizando los elementos qu'hai que char de cuenta pa la entrada de la llingua asturiana nesti área, que tien de pasar necesariamente pela construcción d'un corpus llingüísticu (estaya 4.2).

2 Corpora llingüísticos

2.1 De qué falamos cuando falamos de corpus

Un *corpus llingüísticu* ye una colección de testos orales o escritos d'una llingua, esbillaos en cuenta d'unos criterios llingüísticos esplícitos, pasaos a soporte electrónicu y mínimamente procesaos, y que se manden como amuesa representativa d'ésta pal atropamientu de datos empuestos al so estudiu sistemáticu.[2] Dientro de delles aries dedicaes al estudiu del llinguaxe, l'usu de corpus como recopilaciones de datos pa la xera científica nun ye una práctica nueva. Por exemplu, al sen más axustáu al so étimu llatín, na estaya de los estudios lliterarios dizse-y *corpus* a la totalidá de la obra d'un determináu autor. De la mesma, conozse como corpus el conxuntu de contestos atropaos por lexicógrafos (bien de veces provinentes de la obra de los clásicos), col envís d'iguar y exemplificar les definiciones de los vocablos de la llingua en cuestión.

Estos usos de corpus estrémense, sicasí, de lo qu'anguaño se reconoz como l'oxetu de la *Llingüística de Corpus*. Un corpus llingüísticu estrémase d'un corpus lexicográficu tradicional pol fechu de qu'esti últimu suel recoyer namá amueses fragmentaes de la llingua, anquequier mui infrecuentes y anecdótiques, mientres que l'otru respe por una mayor *representatividá*. Coles mesmes, los corpora llingüísticos estrémense d'otros corpora de datos testuales (como los constituyíos pola obra d'un autor) pola so finalidá: los primeros fáense col envís específicu de sirvir d'encontu p'análisis llingüísticos, y poro diséñense por acio de *criterios llingüísticos esplícitos*; mentes que los segundos entámense a partir de criterios esternos a la llingua de los testos. Otramiente, los corpora llingüísticos caracterícense pol tratamientu col que ye penerada la so información al envís de facilitar la xera d'análisis. Con éses, independientemente del soporte orixinal de los testos que lo componen –oralidá o escritura— los corpora llingüísticos son *informatizaos* y, como se verá, na mayoría de los casos tamién procesaos mínimamente por facilitar l'accesu y la recuperación de la información.

El papel que xuega la dixitalización de la información nos corpora testuales nun ye menor. Cómo ye, que dellos autores refiérense a los corpora llingüísticos actuales como *corpus informatizaos*. Esto tien la so xustificación en contestu históricu de la disciplina. La utilización de colecciones de datos llingüísticos pal analís sistemáticu d'una llingua algamó reconocencia dientro de la tradición de la llingüística de campu a principios del sieglu, mesmo que na escuela estructuralista posterior. Sicasí, esta práctica féxose de menos col surdimientu de los plantegamientos de xeitu mentalista defendíos por Chomsky dende finales de los 50, y que marquen un xiru nos estudios llingüísticos

[2] Esta definición mira d'apautar les distintes caracterizaciones de corpus llingüísticu na bibliografía. Véase, por exemplu Sinclair (1996), Tognini-Bonelli (1996), Torruella y Llisterri (1999) o McEnery y Wilson (2001).

contemporanios. L'oxetivu d'esti nuevu paradigma ye l'estudiu de la gramática universal subxacente a la competencia de los falantes, y coles mesmes, l'actuación contémplase namá como'l reflexu imperfectu del conocimientu llingüísticu. En cuantes que por fuerza reproducen esti nivel d'actuación, los conxuntos de datos llingüísticos atropaos na tradición estructuralista y de la llingüística de campu dexaron d'interesar.

Pero, magar la so impopularidá, los estudios de calter más empíricu nun se desdexen dafechu. Entemás que mientres les décades de los 60 y 70 dieron n'incorporar, como elementu indisociable a esti tipu de planteamientu, la utilización de recursos tecnolóxicos. Con éstes, ente que nos trabayos previos de los estructuralistes y los llingüistes de campu'l soporte material de la información yera'l papel, darréu'l mediu d'almacenaxe vien ser electrónicu y la manipulación de los datos puede beneficiase de procesos automatizaos. Paralelamente, los averamientos de base empirista vuelven pasu ente pasu a ganar adeptos, a cuenta de razones tanto de mena metodolóxica (por casu, aries como l'alquisición del llinguaxe nun son de sofitase n'intuiciones de llingüista), como de formulación teórica de base sobre la naturaleza del llinguaxe natural (una bona introducción a tou esti procesu ye McEnery y Wilson 2001). D'esta miente, nos años 80 el campu conocíu anguaño como la *Llingüística de Corpus* espoxiga gracies a la converxencia d'esa perspectiva adoptada en dellos ámbitos de la llingüística coles posibilidaes de tratamientu de datos que permite'l desenrueldu tecnolóxicu del momentu.

2.2 Esbillando la información: clases de corpora

En cuantes que los corpora llingüísticos tienen como finalidá apurrir información pal estudiu d'una o más llingües, mírase a que constituyan fragmentos representativos d'éstes. La representatividá d'un corpus ye ún de los aspectos sobre los que más se tien aldericao na bibliografía del aria, al rodiu de consideraciones como: qué tamañu tien de tener un corpus que sía daveres representativu; qué clas de testos debe contener; en qué cantidá, etc. Véase d'exemplu Quirk (1992), Biber (1993) o Sinclair (1996).

Mientres l'espoxigue de la llingüística de corpus, nes décades de los 80 y 90, la representatividá plantegóse esencialmente en términos de *tamañu* (según más información, mayor representación de la llingua), y d'*equilibriu* ente los distintos usos de la llingua (variantes dialectales, oralidá versus escritura, diversidá de rexistros y xéneros, etc.). Anguaño, el tamañu sigue valorándose como un elementu importante a la de capturar el mayor númberu de fenómenos posibles d'una llingua. Sicasí, la cata d'una representación permediada de toles variedaes d'usu foi acutándose únicamente al criteriu pa la construcción de los llamaos corpora xenerales (los de reflexar la llingua común na totalidá de les variedaes y ámbitos). La idea de que les aplicaciones finales determinen diseños diferentes de corpus foi cuayando de mou natural, invalidando los plantegamientos qu'imponíen una serie de criterios estándares pal algame de la representatividá.

Veremos les aplicaciones potenciales d'un corpus llingüístico na sección viniente. Pel momentu, ye la d'introducir les posibles clases de corpus, afitaes a cuenta de distintos parámetros complementarios ente ellos:[3]

[3] La mio caracterización ye mínima y malapenes destinada a cubrir de mou básicu la llaguna de que pueda cadecer un lector non familiarizáu col tema. Pa una clasificación más refecha ye de vese Torruella y Llisterri (1999).

- **Soporte orixinal de los datos:** tenemos *corpus escritos* o *corpus orales*, según tean constituyíos por testos escritos o orales.
- **Llingües que lu constituin:** pueden estremase *corpora monollingües* y *multillingües*. Estos últimos denómense *corpus paralelos* a tar constituyíos polos mesmos testos en caúna de les llingües representaes.
- **Variación llingüística que se mira a representar:** diatópica, diastrática y/o diacrónica. Dellos corpora representen má una de les variedaes de la llingua. A otra mano tan los *corpora xenerales*, que miren de representar varies d'elles, xeneralmente acutaes cronolóxicamente. Bona cuenta, nesti tipu de corpus ye mester asegurar l'equilibriu de representación ente les distintes variedaes.
- **Nivel d'especialización de los testos:** los *corpus especializaos* concéntrense en testos de determinaes aries d'especialidá. La finalidá suel ser mui específica, xeneralmente dientro del ámbitu de la terminoloxía y la terminografía.
- **Criterios d'esbilla de los datos:** estrémense corpora constituyíos a partir de testos enteros y los que contienen namá fragmentos (xeneralmente de llargor constante), que se conocen como c*orpora d'amueses*. Anguaño, esta dixebra tira a escaecese por mor de los los problemes que planteguen los corpora d'amueses pa la recuperación del contestu global d' usu de les espresiones llingüístiques.
- **Criterios d'actualización de los datos:** dacuando interesa la renovación periódica de los materiales d'un corpus. Esto ye sobremanera en corpora que respen por representar la llingua actual, nos que'l fragmentu de testos más antiguos ye regularmente sustituyíu por otru de materiales recientes. Esti tipu de corpora conozse como *corpus monitor*.

2.3 Utilizando la información: aplicaciones posibles d'un corpus

Les utilidaes qu'ufierten los corpora llingüísticos beneficien a distintos campos y disciplines rellacionaes col estudiu del llinguaxe. Pueden xuntase en tres grandes niveles: dende les ciencies del llinguaxe, un corpus val de base pa estudios de plantegamientu teóricu. Por casu, d'abastecedor de contestos pa la validación de modelos de descripción de la llingua, pal analís de la rellación ente dos llingües (a partir de corpora paralelos), o bien como indicador de la variación y les tendencies d'usu, d'interés n'estayes como la dialectoloxía y la sociollingüística, y qu'en casu de les llingües minorizaes puede emponese a determinaes decisiones de planificación llingüística. De la mesma, un corpus puede utilizase pal desenrueldu de productos de base llingüística. Fai'l casu de la ellaboración o meyora de gramátiques y diccionarios (esbilla del léxicu más frecuente, usu d'exemplos reales, distinción de sentíos a cuenta de los datos y non de la intuición del lexicógrafu, etc.); na creación de vocabularios especializaos; o na construcción de ferramientes pal aprendizaxe d' una llingua estranxera.

Nun segundu nivel, los corpora mándense tamién nel desenrueldu de tecnoloxía llingüística. Por exemplu, los corpora paralelos úsense pa entrenar sistemes de traducción automática, los corpora orales nel ámbitu de reconocencia y síntesis de fala, y los testuales pal entrenamientu de correctores ortográficos y gramaticales.

Finalmente, na estaya del procesamientu del llinguaxe natural los corpora utilícense tamién de manera retro-alimentativa, na meyora de les ferraminetes de procesamientu de corpus (sobremanera, pal entrenamientu de les de base estadística), talos como

analizadores morfolóxicos y sintácticos, o etiquetadores semánticos. Vemos darréu cuáles son estes ferramientes y el so valir.

2.4 Tratando la información: corpora anotaos

Los corpora llingüísticos nun son solo esto: un conxuntu de testos caracterizaos a cuenta de determinaos criterios, que s'embalaguen en soporte electrónicu, y que son utilizaos con una finalidá específica. Precisamente a cuenta de la so finalidá como fonte d'abastecimientu de datos, ye mester que la información que contién sía bona d'algamar, y néstes la dixitalización de testos nun ye abondo. Poro, na mayoría de casos los corpus llingüísticos presenten, amás del so conteníu testual orixinal, un nivel metallingüísticu d'información qu'esplicita les característiques gramaticales de los distintos niveles de constituyentes: dende traces morfolóxiques de los elementos léxicos hasta, potencialmente, característiques de los párrafos o unidaes testuales mayores, en casu de corpus escritos; nos corpus orales, dende información fonética hasta estructura prosódica. A cuenta de la mio esperiencia personal dientro de la Llingüística de Corpus, nesta estaya y lo que sigue del artículu voi concentrame casi dafechu na parte de corpus escritos. El llector, sicasí, afayará abondes referencies bibliográfiques a la d'enanchar la so conocencia tocantes a corpus orales. Véase por casu Llisterri (1997, 1999).

2.4.1 Niveles d'anotación

Los corpora llingüísticos qu'apurren esti nivel de metainformación suelen referise como *corpora etiquetaos*, cuidao que la información codifícase en testu pentemedies de códigos específicos o conxuntos d'etiquetes (o *etiquetarios*). De la mesma, tamién puede dicise *corpora anotados*.[4] Per un sitiu, l'etiquetaxe d'un corpus facilita la xera de remanar y recuperar información. Por casu, pa un lexicógrafu o un gramáticu que quiera analizar los contestos d'usu d'un verbu determináu, disponer de toles formes d'esti verbu identificaes pola so rellación con un únicu lema quita de tener que buscar los contestos d'usu a partir de la enumberación refecha de toles formes posibles. Otramiente, l'etiquetaxe permite una cuantificación de datos más sofisticada que la pura frecuencia d'usu de cada forma léxica (por exemplu, distribución de tiempos verbales, clases de complementos pa determinaos predicaos, tipos de modificación aplicada a una clas de nomes particular, coocurrencies frecuentes d'elementos léxicos, variantes dialectales, etc.). Con éstes, pueden albidrase tendencies de comportamientu llingüísticu dientro del fragmentu de llingua analizáu.

L'etiquetaxe d'un corpus suel informar sobre distintos niveles: l'*etiquetaxe estructural*, por casu, marca la organización estructural del testu: títulu, subtítulu, capítulu, párrafu, sección, subsección, etc., hasta frase, xeneralmente. L'*etiquetaje morfolóxicu* aplícase al nivel de los elementos léxicos, indicando la so categoría gramatical y, en llingües como les romances, tamién la información de flexón verbal y

[4] Nos últimos años, l'usu de corpus non etiquetaos féxose tamién una práctica relativamente avezada, pola economía de recursos y tiempo que conlleva. Sicasí, na mayoría de casos trátase de corpus utilizaos pal entrenamientu de ferramientes estadístiques de procesamientu del llinguaxe (veremos más alantre a qué me refiero), envede corpus pensaos pal desenrueldu de preseos de base llingüística (diccionarios, gramátiques) o pa la descripción de la llingua. Esti trabayu nun va aportunar nesa perspectiva.

nominal. Xeneralmente, sobre esti nivel aplícase l'*etiquetaxe sintácticu*, que puede realizase con más o menos fondura, d'acordies cola aplicación pa la que se plantega'l corpus. Con éstes, l'anotación puede ser superficial, de simples grupos nominales y/o verbales, o más complexa, indicando les funciones sintáctiques de los distintos componentes d'una fras (suxetu, oxetu, modificadores, etc.).[5]

Éstos son en grandes traces los niveles d'anotación más comunes. Por embargu, otros niveles d'información pueden faese igualmente prestosos d'acordies cola aplicación específica a la que s'empobine'l corpus. Por exemplu, puede introducise tamién anotación de calter semánticu o pragmáticu. Pero, según nos niveles anteriores podía trabayase dende un plantegamientu enforma neutru tocantes a cualaquier teoría llingüística, dende'l nivel semántico l'etiquetáu mira o a basase nuna visión particular de la semántica o la pragmática, o a especializase nun aspectu concretu, como les rellaciones anafóriques, les unidaes d'información temporal, los marcadores discursivos, los papeles temáticos de los predicaos verbales, etc.[6]

2.4.2 Ferramientes de procesamientu de corpus

Toos estos niveles d'etiquetaxe faense al traviés de varies capes de procesamientu automatizáu. Por exemplu, pa l'anotación morfolóxica empléguense los programes conocíos como *analizadores morfolóxicos*, que trabayen cola collaboración d'un *diccionariu* en formatu electrónicu nel que cada elementu léxicu xúncese a una o más categoríes gramaticales y, en llingües como les romances, d'un *lematizador*, que ye'l d'encargase de dixebrar nes palabres la raíz y les terminaciones. Na xera d'etiquetaxe sintácticu úsense los llamaos *analizadores sintácticos*, que como ya se comentó, pueden trabayar a un nivel más o menos fondu d'analís: dende la simple detección de grupos nominales y verbales, hasta'l marcaxe de funciones y rellaciones de dependencia. Pa l'anotación del nivel semánticu (y mesmo'l pragmáticu), nun hai ferramientes específiques que permitan l'agrupación baxo un únicu denomador, según asocedía colos niveles previos. Sicasí, vaga la pena comentar la esistencia de los *etiquetadores semánticos*, sobre los que se tien trabayao enforma apocayá. Trátase de ferramientes qu'asignen etiquetes semántiques a les unidaes léxiques a partir de la so clasificación nuna ontoloxía determinada, lo que, naturalmente, quiere un llabor previu de desambiguación léxica.

---

[5] Una y bones enforma bibliografía d'esti campu vien n'inglés, cuento importante introducir equí la terminoloxía de los distintos niveles d'etiquetaxe tamién nesta llingua. A la fase d'anotación morfolóxica conózse-y como *part-of-speech (POS) tagging*. Al nivel d'etiquetaxe sintácticu, como *parsing*, identificando tamién el marcaxe sintácticu superficial como *shallow parsing* o bien *chunking*.

[6] Exemplos particulares de corpus nesti grupu son: el trabayu presentáu en Fligelstone (1992), que se caracteriza pola anotación de rellaciones anafóriques; el *RST Corpus* (Carlson *et al*., 2003), nel que los testos anótense a partir de la teoría de discursu *Rethorical Structure Theory* (Mann y Thompson, 1988); el *TimeBank* (Pustejovsky *et al*. 2003), nel que namá s'anoten les eventividaes y les expresiones temporales col envís de mandase d'elles n'estudios sobre razonamientu temporal; o tamién el Corpus de Operaciones Metalingüísticas Explícitas (Rodríguez, 2003), nel que s'anotaren les operaciones de creación de conocimientu nel marcu del discursu científicu.

D'últimes, ye importante mentar tamién les *ferramientes pa la esplotación de corpus*, que son básiques na investigación lexicográfica y llingüística. Les más destacaes equí son les que realicen busques de (secuencies de) palabres, presentando los resultaos ordenaos alfabéticamente y col contextu oracional, y ufiertando arriendes la posibilidá de cómputu de frecuencies.

Les ferramientes de procesamientu de corpus estrémense en xeneral según el plantegamientu de base de que partan: d'uno, tán les ferramientes de base simbólica; d'otru, les de base estadística. Si ye les primeres, desenvuélvense a partir de conocimientu llingüísticu, y poro, tan en contactu con delles de les vertientes de la llingüística teórica. Les segundes básense en modelizaciones estocástiques del fragmentu de llingua col que se trabaya. El conocimientu llingüísticu utilizáu nesti segundu casu ye mínimu, si non inesistente, y poro en munchos casos nun dexa, desque desenroldada l'aplicación en cuestión, faer una abstracción en términos llingüísticos de los fenómenos trataos. Sicasí, suel dar resultaos meyores que'l de una ferramienta de base simbólica desenroldada per un períodu de tiempu equivalente.

Arriendes de l'averamientu simbólicu o estadísticu al problema, el campu dedicáu al desenvolvimientu de toes estes ferramientes pal tratamientu de testos escritos conozse como *Procesamientu del Llinguaxe Natural* (PLN) o, en dellos contestos tamién, *Inxeniería Llingüística*; mentanto que no que fai al procesamientu de testos orales fálase de *Tratamientu de la Fala*. Equí voi emplegar el términu xenéricu de *tecnoloxíes del llinguaxe* (o'l ximielgu *tecnoloxíes llingüístiques*), a la de referime globalmente a estes dos aries en xunto a la Llingüística de Corpus. Al dicir del so nome, el so denomador común ye l'usu de tecnoloxía pal tratamientu del llinguaxe.

3 Esperiencies de referencia

Desque afitaos los preliminares, esta estaya presenta en grandes traces les esperiencies previes de proyectos de corpus que cuido de más interés pa lo qu'equí nos lleva'l tiempu: la construcción d'un corpus llingüísticu pa la llingua asturiana.

Les primeres esperiencies que derrompen la nueva dómina na Llingüística de Corpus danse pa la llingua inglesa. Anguaño, constituin obres de referencia y, magar que tean bonamente documentaes na bibliografía de la disciplina, vaga bien presentales mínimamente. La so naturaleza de finxu ye claramente resultanza del calter pioneru de la investigación en delles comunidaes onde se fala esta llingua. Esisten, bona cuenta, corpus n'otres llingües, cuando cercanes xeográficamente (como el francés o l'alemán), cuando más alloñaes: dende'l xaponés, el coreanu y el mandarín, hasta les recientemente tan preciaes llingües falaes en mundu musulmán, como l'árabe y el farsi.[7] Cuento, sicasí, que queda fora del algame d'esti trabayu faer un repás, anquequier mínimu, de los casos más relevantes pa cada llingua.

[7] A esti sen, vemos que sigue siendo parte de la comunidá de fala inglesa la de decidir promocionar recursos nuna o otra llingua. Ye indicativo la creciente demanda de llingüistes, traductores ya intérpretes en delles d'estes últimes llingües xenerada de magar el 11 de setiembre del 2001. L'emburrión a la investigación en delles llingües vese afaláu por intereses de la comunidá de les axencies d'espionaxe ya intelixencia militar, y con oxetivos que poco o nada tienen que ver cola investigación puramente científica o l'apreciu de valores culturales.

Otra manera, voi revisar tamién los proyectos esistentes dientro del marcu xeográficu de la Península Ibérica, centrándome dafechu naquellos que se desenvuelven al abellu del mesmu contestu políticu nel que s'inclui l'asturiano; esto ye, l'Estáu Español (sección 3.2).

3.1 Los pioneros

Arriendes de la mesma evolución del campu, nos corpus del inglés son d'estremase dos xeneraciones. La primera inclui los corpus entamaos nes décades de los 50 y 60, de la que los averamientos empíricos al llinguaxe dexaron d'interesar, particularmente dientro de la llingüística desenroldada nel contestu de fala inglesa. Los corpus d'esti períodu inicial caracterícense por ser de tamañu modestu en comparanza con dellos de los corpus construyíos de más recién. Los más destacables son:

- *Survey of English Usage*:
  Corpus d'inglés británico entamáu en 1955 y desenroldáu a lo llargo de, al aldu, 30 años, nel University College London. Contién un millón de palabres, ente testos orales y escritos. Ye de remarcase que los materiales d'esti corpus emplegáronse como base pa la constitución de la conocida gramática de referencia del inglés (Quirk *et al.*, 1985).

- *Brown Corpus*, *Lancaster-Oslo/Bergen Corpus* (LOB) y *Kolhapur Corpus*:
  Corpus de testos escritos nes variantes d'inglés americano, británico ya inglés de la India, respectivamente. El primeru atropóse na década de los 60 y los otros dos n'etapes posteriores, como corpus equivalentes al anterior pa otres variantes del inglés. Contienen toos un millón de palabres y presenten los testos agrupaos en 15 categoríes diferentes, procurando d'esta miente una bona representatividá de los distintos usos de llinguaxe escrito. Son, poro, corpus con vocación de referencia (Francis y Kucera, 1964).

- *London-Lund Corpus*:
  Corpus oral d'inglés británico, de 500.000 palabres. Contién esclusivamente trescripciones de material falao, orixinario de dos proyectos diferentes: de la parte oral del *Survey of English Usage* presentáu enantes, y del *Survey of Spoken English*, entamáu na Lund University en 1975 como proyectu hermanu del anterior (Svartvik, 1990).

Anguaño, la dimensión de los corpus tien variao enforma por mor de los avances tecnolóxicos. La capacidá d'atropamientu ye bramente mayor de lo que yera de la qu'empezara la tecnoloxía dixital, y los procesos de tratamientu de datos son abondo más rápidos. A casi 50 años de magar la creación del primer corpus modernu, los corpus qu'hai disponibles pal inglés son munchos y d'estremada mena. D'ente ellos interesa mentar los vinientes pol so calter de referencia dientro del aria y la so trescendencia na estaya de la llingüística aplicada:

- *Bank of English*:

Entamáu en 1991 de parte de la Birmingham University y COBUILD, una división de la editorial Harper Collins. El so principal envís ye mandase como fonte de datos n'estudios llingüísticos y la fechura de diccionarios. Trátase d'un corpus monitor; esto ye, un corpus al que se-y amiesta dacuando material nuevo. La so última edición, de 2002, atropa 450 millones de palabres (Sinclair, 1987).[8]

- *British National Corpus* (BNC):
  Contién un total de 100 millones de palabres del inglés moderno, tanto del rexistru escritu como del oral, qu'inclui hasta parolaes coloquiales. Foi atropáu de parte d'un xermandía formada por editores británicos, instituciones académiques como la Oxford University y la Lancaster University. Tanto nesti casu como nel anterior, el so diseñu respe pola mayor representatividá posible (Aston y Burnard, 1998).[9]

3.2 La llingüística de corpus nel nuesu ámbitu xeográficu

El desenrueldu de corpus llingüísticos, mesmo qu'otres ferramientes pal tratamientu del llinguaxe natural nel ámbitu del Estáu Español, reproduz la división ente llingües oficiales o co-oficiales, d'uno, y llingües ensin reconocencia oficial, d'otro. En mayor o menor grau, el primer grupu de llingües (euskera, catalán, español y gallegu) dispón anguaño de presea y recursos de procesamientu del llinguaxe, grupos d'investigación más o menos afitaos aplicaos al so desenvolvimientu, y (non por último, menos importante) la voluntá institucional de sofitar esta llinia de trabayu. Sicasí, la Llingüística de Corpus y el Procesamientu del Llinguaxe Natural son tierres entá non derrotes en casu de llingües ensin oficialidá como l'asturianu.

El Procesamientu del Llinguaxe Natural nel nuesu ámbitu surde metá los 80, al rodiu de los proyectos de traducción automática que s'entamen en Barcelona y Madrid, tanto dende empreses privaes como financiaos públicamente pola Comisión Europea. El campu pasa daquella per una dómina d'euforia paralela a la espectación qu'esiste nel marcu européu más xeneral, pero, desque nos 90, esos proyectos dan n'abandonase porque la probeza de los resultaos obtenidos hasta entós, en comparanza colos recursos invertíos, apunta un fracasu nos plantegamientos de base.[10]

Sicasí, de la qu'estos proyectos se desmantelen, yá hai la infraestuctura creada en términos d'instituciones y recursos humanos a la de siguir la investigación empobinada agora a la constitución de recursos llingüísticos esenciales (como corpus, diccionarios en formatu electrónicu y bases de datos léxicos), y a la igua de ferramientes básiques pal procesamientu del llinguaxe. A esto tien que se-y amestar el fechu de que, al par desendolcu de proyectos de traducción automática na década de los 80, yá esiste dalguna institución trabayando na estaya de la Llingüística de Corpus, como fai'l casu del proyectu del *Diccionari del Català Contemporani* col so *Corpus Textual Informatitzat* (descritu darréu). Coles mesmes, ye metanos los 90 de la que la Llingüística de Corpus espoxiga nel nuesu ámbitu xeográficu, cuayando tanto en contestu español y en Cataluña (onde surdieren los primeros enteinos en traducción automática), como en País Vascu y

[8] Véase tamién http://titania.cobuild.collins.co.uk/boe_info.html.

[9] Véase tamién: http://www.natcorp.ox.ac.uk.

[10] Véase Abaitua (1999) pa una refecha caracterización d'esti períodu.

Galicia, col desenrueldu consiguiente de recursos para caúna de les llingües falaes nestes zones.

Ufierto darréu una amuesa representativa, anque non refecha, de los corpus llingüísticos esistentes anguaño d'estes cuatro llingües: euskera, catalán, español y gallegu. Bien que nenguna d'elles presenta un panorama asemeyáu al del inglés, tampoco nun se caractericen pola situación llaceriosa na que s'afayen les llingües ensin reconocencia oficial. De la mesma, albídrase daqué disparidá nel nivel de desenrueldu del aria en caúna de les llingües. Ello correspuende en llinies xenerales colos datos que s'apurren na parte dedicada al fomentu y desenvolvimientu del aria dientro'l plan nacional d'investigación científica del gobiernu vascu[11]. Con encontu nun estudiu recién ellaboráu de parte'l Consorciu Européu EUROMAP, ésti detalla que'l repartu d'esfuerzu d'investigación y desenvolvimientu empobinada en caúna de les llingües en cuestión en marcu conxuntu del estáu ye ésti: 60 % pal español, 20% pal catalán, 8,5% pal gallegu y finalmente 7% pal euskera. Sigo esti orde de llingua de más a menos recursos na presentación de los proyectos de corpus en caúna d'elles. Sicasí, voi centrame más nos proyectos desenroldaos pa les llingües co-oficiales que nos del español: anque en menor grau, parten col asturianu la so situación minoritaria y minorizada y, poro, la so esperiencia dientro de la Llingüística de Corpus resulta de gran utilidá a la d'apuntar un proyectu nesti terrén, con recursos y sofitu institucional acutáu.

3.2.1 Corpus d'español

Esisten dellos corpus pa esta llingua. Dos de los más destacaos son los corpus de referencia atropaos pol Instituto de Lexicografía de la Real Academia Española, ún diacrónicu y otru del español actual. Dos referencies que detallen el desenrueldu d'estos proyectos son Pino *et al.* (1999) y Sánchez-León *et al.* (1999). Dambos corpus incluin documentos escritos en toles variantes, tanto peninsulares como estrapeninsulares, y tán eminentemente empobinaos a mandase como recursu básicu nel trabayu lexicográficu d'esta institución:

- *Corpus Diacrónico del Español* (CORDE)
  Corpus que mira de representar la variación del español a lo llargo de la so historia escrita. Contién documentos de tres dómines básiques: Edá Media, Sieglos d'Oro y Época Contemporania (de mano, hasta 1975). Anguaño presenta un total de 145 millones de palabres y puede consultase per internet.[12]

- *Corpus de Referencia del Español Actual* (CREA)
  Corpus monitor, constituyíu de testos representativos de les distintes variantes del español actual. Compriende un trechu de 25 años, que s'actualiza de xemes con materiales más recientes. Esto significa que los documentos más vieyos treslládense al corpus CORDE. Anguaño contién 145 millones de palabres.

---

[11] *Plan Nacional de Investigación Científica, Desarrollo e Innovación (2000-2003).* 'Propuesta de acción estratégica en el área sectorial "Sociedad de la Información": Industria de la Lengua e Ingeniería Lingüística'. Eusko Jaurlaritza (Gobierno Vasco). http://www.euskadi.net/euskara/datos/azkeninforme.pdf (26/09/2003).

[12] http://www.rae.es/cordenet.html

Esisten otros corpus del español entamos por editoriales, como por exemplu:

- *Corpus CUMBRE*
  Desenroldáu pola Editorial SGEL con fines lexicográficos. Atropa 20 millones de palabres, vinientes de testos tanto escritos como orales del español peninsular ya hispanoamericano. Los testos escritos pertenecen a distintos xéneros (lliterariu, periodísticu, ensayu) y toquen distintes estayes d'especialidá (filosofía, historia, ciencia, derechu, economía, etc.). Los testos orales proceden de grabaciones de programes de radio y televisión. Según les amueses orales son de la década de los 90, les escrites algamen el períodu ente los 50 y los 90 (Sánchez *et al.* 1995).

De la mesma, Vox Bibliograf trabaya col so corpus d'al rodiu de 10 millones de palabres, y la editorial SM con otru de 60.000 palabres.

Amás d'estos corpus, esisten otros de menores dimensiones, cuando de llingua xeneral pero con un envís específicu (ye la de los corpus d'editoriales que vienen de mentase), cuando de llinguaxe acutáu, como'l *Corpus de vocabulario del niño de 6 a 14 años* de la Universidá de Granada, o bien el *Corpus 92, Lengua escrita por aspirantes a estudios universitarios* de la Universitat Pompeu Fabra.

D'últimes, vaga tamién mencionar la esistencia de corpus puramente orales, creaos pal desendolcu d'aplicaciones na estaya de les Tecnoloxíes de la Fala. Como yá anuncié enantes, nun voi centrame nellos.

3.2.2 Corpus de catalán

Los corpus de mayor envergadura detállense darréu. Información adicional sobre corpus testuales nesta llingua pueden afayase en Soler i Bou (1998) y, a nivel más xeneral, Llisterri (2000) ufierta una introducción abondo completa sobro los recursos y ferramientes de procesamientu del llinguaxe esistentes anguaño.

- *Corpus Textual Informatitzat de la Llengua Catalana (CTILC)*
  Corpus de testos escritos de rexistru variáu (lliterariu y non lliterariu) desendolcáu nel Institut d'Estudis Catalans como componente básicu na creación del *Diccionari del Català Contemporani* (DCC). Trátase, poro, d'un corpus de llingua xeneral, que respe pola representatividá y que foi diseñáu col envís de mandase como base na investigación lexicográfica. Compónenlu al aldu de 4000 testos dataos ente 1832 (fecha simbólica del aniciu de la dómina conocida como la *Renaixença* de la llingua catalana, cola publicación de la *Oda a la pàtria*, de Bonaventura Carles Aribau) y 1988. En total, recueye 52,3 millones de palabres. El corpus ta lematizado y etiquetáu morfosintácticamente (Rafel 1992-1993, 1996). Amás, compleméntase con una base de datos lexicográfica constituyida a partir de 13 diccionarios espublizaos en mesmu períodu, y sirvió yá pa la creación del *Diccionari de Freqüències* (Rafel i Fontanals, 1996-1998).

- *Corpus del Català Contemporani de la Universitat de Barcelona (CUB)*
  Corpus de testos escritos y orales, estructuráu en siete subcorpus independientes que se rellacionen ente ellos pol envís de configurar conxuntamente una caracterización del catalán actual dende la so variación diatópica, diastrática y diafásica. Contién: un

subcorpus de variedaes xeográfiques (*Corpus Oral Dialectal*); ún de variedaes sociales (*Corpus Oral Social*); y los cinco restantes de variedaes funcionales (*Corpus Oral de Conversa Col·loquial*, *Corpus Oral de Registres, Corpus Oral de Publicitat, Corpus d'Informatius Orals, Corpus Escrit de Català Actual*). Los subcorpus escritos tán lematizados y anotaos morfosintácticamente. Los subcorpus orales tán transcritos y, según la naturaleza de los datos, presenten anotación discursiva, de grupos tonales, o bien lematización y etiquetáu morfosintácticu. Si en casu del CTILC, la utilidá que se quier ye la investigación lexicográfica, nésti mírase a crear una base pa estudios específicos sobre la variación del catalán. Bones introducciones al proyectu son Boix (1996) y Viaplana (2000).

- *Corpus textual plurilingüe especialitzat*
  Corpus ellaboráu nel Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra (Bach *et al*. 1997). Contién testos escritos en catalán, español, francés, inglés y alemán, pertenecientes a las siguientes aries d'especialidá: mediu ambiente, informática, medicina, derechu y economía. La parte del catalán cuenta con unos 5 millones de palabres, lematizaes y con etiquetáu morfosintácticu.

Otros proyectos de corpus del catalán son: el *Corpus Parole*, de 21 millones de palabres desendolcáu en marcu d'un proyectu européu, y el *Corpus de Diatopia Diacrònica de la Llengua Catalana*, afaláu dende la Universitat de Barcelona. De la mesma, esisten dellos corpus de fala atropaos específicamente pal desenvolvimientu d'aplicaciones de tecnoloxíes de la fala.

3.2.3 Corpus de gallegu

Una panorámica xeneral de la situación del gallegu ufiértala García-Mateo *et al.* (1998). Equí interesa destacar los corpus siguientes, toos ellos en tenores:

- *Corpus De Referencia Do Galego Actual* (CORGA)
  Esti corpus ye ún de los proyectos entamaos en Centro Ramón Piñeiro para a Investigación en Humanidades, institución creada de parte de la Xunta de Galicia por afalar la enseñanza, la investigación y l'usu de la llingua gallega. El CORGA plantégase como un corpus de referencia de testos escritos y orales, producíos ente 1975 y 2004 (fecha prevista pa la so finalización), y mira d'abarcar diferentes rexistros del gallego actual. Anguaño embalaga unos 17,5 millones de palabres anque aspírase un total de 25 millones, anotaes cola categoría morfosintáctica. Los detalles del proyectu, mesmo que la última versión de CORGA, emitida en xineru del 2003, puede consultase per internet.[13]

- *Corpus do galego moderno* (CGM)
  Corpus atropáu nel Instituto da Lingua Galega. Contién al rodiu de 10 millones de palabres, vinientes de testos escritos dende'l S.XVII hasta güei. Mira d'incluyise ende testos de la lliteratura oral, histories, refranes, canciones populares, etc. L'envís principal del corpus ye lexicográficu y pal estudiu de la llingua gallega en xeneral.

[13] http://corpus.cirp.es/corga/

- *Arquivo do Galego Oral*
  Corpus oral, igualmente atropáu pol Instituto da Lingua Galega. Estrémase en dos subcorpus: l' *Arquivo do galego popular* y l' *Arquivo do galego culto*. El primeru contién más de 1000 hores de grabaciones orales efectuaes ente 1974 y 1998, ya inclúi fablantes de distintes edaes y variedaes dialectales. El segundu inclúi grabaciones de charles, conferencies, y meses redondes sobre temes del ámbitu politicu y social.

- *Corpus Linguístico da Universidade de Vigo* (CLUVI)
  El CLUVI desendólcase en *Seminário de Linguística Informática* de la Universidade de Vigo (Aguirre Moreno *et al.* 2001, 2002, 2003). Trátase d'un corpus escritu y oral que recueye testos contemporanios de distintos rexistros especializaos: xurídico-alministrativu, periodísticu, informáticu y lliterariu. Otra manera, los testos pueden ser monollingües en gallegu, traducciones gallegu-español y traducciones inglés-gallegu. La parte de testos d'aniciu escritu estrémase en cuatro subcorpus d'al rodiu d'un millón de palabres caún: el corpus paralelu TECTRA, constituyíu de testos de rexistru lliterariu inglés-gallegu; el corpus paralelu LEGA, de testos xurídico-administrativos gallegu-español; el corpus monollingüe XIGA, de testos del ámbitu informáticu escritos en gallegu; y finalmente el corpus monollingüe MEGA, del ámbitu de los medios de comunicación social. De la mesma, trabáyase na contrucción d'otra parte adicional constituyida de testos paralelos portugués-gallegu. Quier llegase al nivel d'anotación morfosintáctica, mesmo qu'a la allineación oracional de los subcorpus paralelos. Les aplicaciones que se quieren construyir en base a esti corpus son varies: dende la construcción de ferramientes básiques como estractores d'información léxica, terminolóxica y fraseolóxica, hasta aplicaciones más sofisticaes y de llargu plazu, como la estracción d'información, la sumarización o la recuperación d'información *on-line*.

Ye de remarcase tamién la esistencia d'otru corpus oral, VOGATEL, desenroldáu de parte de la empresa Telefónica I+D en collaboración cola Universidade de Vigo. El so envís ye puramente l'entrenamientu de ferramientes pal tratamientu de la fala.

3.2.4 Corpus d'euskera

L'euskera tien dos corpus llingüísticos principales, dambos de testos escritos. Preséntense n'Agirre *et al.* (1998), dientro del marcu xeneral pal procesamientu del euskera que se desenvuelve na Euskal Herriko Unibertsitatea. Esisten amás otros corpus de calter menor, confeccionaos de parte de grupos d'investigación por entrenar les ferramientes de base estadística que construin. Equí voi centrame má nos corpus de referencia, que son:

- *Orotariko Euskal Hiztegia* (OEH, 'Diccionariu Xeneral Vascu'):
  Corpus testual diacrónicu que se manda como base na creación del diccionariu descriptivu d'Euskaltzaindia, l'academia de la llingua vasca. Atropa un total de 5.800.000 palabres, vinientes de 310 obres escrites en distintes variedaes del euskera, dende'l S.XVI hasta la d'entamase la estandarización de la llingua, contra 1960.

- *XX. mendeko Euskararen Corpus Estatistikoa* (‘Corpus Estadístico del Euskera del S.XX’)
  Corpus del euskera actual, constituyíu por amueses de testos vinientes d’al rodiu de 6000 obres escrites mientres el sieglu XX. En ficies de recoyer la máxima variedá léxica y estructural posible, estableciose en base a una esbilla aleatoria del inventariu d’obres escrites n’euskera, atropáu por UZEI (*Terminologia eta Lexikografiako Zentroa*, Centru Vascu de Terminoloxía y Lexicografía). Estes obres pertenencen a toa triba de xéneros, dende poesía, teatru y lliteratura infantil, hasta investigación o llibros de testo. Amás, miróse a representar caún de los cuatro períodos relevantes de la historia del euskera del sieglu XX: 1900-1939 (dende l’entamu de sieglu hasta les guerres), 1940-1968 (de magar la posguerra a la nacencia del euskera estándar), 1969-1990 (dende los cambios producíos pol estándar hasta les publicaciones de les normes de Euskaltzaindia), 1991-1999 (posterior a la normativa). El corpus atropa un total de 4,7 millones de palabres, lematizaes col lema estándar y el de la variante dialectal, y etiquetaes cola categoría gramatical y les distintes acepciones. Diseñóse col envís de representar les distintes variedaes diatópiques y diafásiques del euskera contemporanio. Poro, ta pensáu como fonte de datos qu’espeye l’usu real de la llingua, tornándose d’una función puramente prescriptiva.[14]

4 Un proyectu de corpus pal asturianu

4.1 Tecnoloxía llingüística y llingües minoritaries

Un proyectu de corpus llingüísticu pal asturianu tien de char cuenta necesariamente del so estatus de llingua minoritaria. En cierto, les posibilidaes de desendolcu de tecnoloxía llingüística pa una llingua de complexón sana son bien superiores a les que tien una llingua como l’asturianu o l’euskera. El problema básicu de les llingües identificaes como minoritaries ye que’l nivel de recursos de que disponen, materiales y humanos, ye enforma menor que’l de les mayoritaries. A ello tien que se-y amestar el ruin, si non nulu, interés comercial que presenten, lo que supón un costu mui altu pa la investigación y el desenvolvimientu nel ámbitu.

El problema nun ye, sicasí, esclusivu de les llingües minoritaries. Hai llingües mayoritaries que s’afayen na mesma situación, como por casu l’hindi, magar los sos 180 millones de falantes (Somers 2001). Amás, la etiqueta de llingua minoritaria presenta della ambigüedá. Si por minoritaria entendemos llingua de pocos falantes, debemos incluyir llingües como’l finladés (4,7 millones en Finlandia) o’l suecu (7,8 millones en Suecia)[15]. Nestos casos, por embargu, la condición minoritaria nun quita un bon allugamientu nel ránquing de llingües pa les qu’esisten aplicaciones pal procesamientu del llinguaxe: el finlandés alcuéntrase na 6ª posición y el suecu na 7ª, darréu del inglés, l’alemán, el francés, l’español y l’italiano, d’acordies el Natural Language Resource

[14] Na páxina d’UZEI (http://www.uzei.com/default_cas.html) ufiértase más información. Otra manera, el corpus ye consultable per internet (http://www.euskaracorpusa.net/XXmendea/Konts_arrunta_fr.html).

[15] Datos vinientes de: http://www.ethnologue.com.

Registry (NLRS).[16] Per otru sitiu, si de llingua minoritaria entendemos llingua en situación minoritaria dientro del marcu políticu nel que s'afaya, el términu yera d'aplicase tamién a les llingües d'inmigrantes con rellación al so país d'adopción, anque sían de llingües mayoritaries nel so país d'aniciu.

Ye mester entós considerar cuálos son los elementos que garanticen o torguen l'accesu d'una llingua a les tecnoloxíes del llinguaxe. Veremos darréu como l'estatus de cada llingua determina de mou xeneral la rellación qu'ésta caltién coles tecnoloxíes del llinguaxe. Finalmente, identificaremos na sección siguiente les traces que caractericen la llingua asturiana y aventuraremos una estratexa posible pa la so entrada en campu de la tecnoloxía llingüística

Los factores qu'entren en xuegu nel accesu d'una llingua a les tecnoloxías del llinguaxe son esencialmente les siguientes:

- **Númberu de falantes:** a mayor númberu falantes, mayor ye la rentabilidá económica que supón el desendolcu de tecnoloxía llingüística y, poro, l'interés comercial que ye a movilizar la inversión nel aria de parte d'empreses privaes. L'exemplu paradigmáticu de rentabilidá comercial ye, de xuru, l'inglés.

- **Sofitu institucional:** Nos casos nos que nun hai una rentabilidá económica visible, el sofitu y promoción de la llingua de parte de l'administración son básicos a la de garantizar el desenvolvimientu d'aplicaciones de tecnoloxía llingüística. Naturalmente, el sofitu dende'l gobiernu facilita tamién el d'otres instituciones, como fundaciones culturales privaes con capacidá de mecenalgu. Nesta situación alcuéntrense llingües de pocos falantes, como'l finlandés o l'holandés.

- **Diglosia:** La entrada d'una llingua en campu tecnolóxicu depende tamién del grau nel qu'ésta sía emplegada en tolos ámbitos y rexistros d'uso, orales como escritos. Nuna situación de diglosia, una llingua supeditada y arrequexada dafechu a la comunicación oral dientro d'ámbitos familiares va tener bien difícil la creación de tecnoloxía llingüística: per un sitiu porque los recursos esistentes como puntu de partida (dende diccionarios y gramátiques hasta un cuerpu de testos, escritos o grabaos, abondo voluminosu) van ser mui escasos; pel otru, porque'l desendolcu d'aplicaciones va pasar prioritariamente pela llingua subordinante. Con éses, la situación d'escasez de tecnoloxía llingüística na que s'atopa l'hindi, magar el so altu númberu de falantes, seique ye debida a la so supeditación al inglés.

- **Accesu a les tecnoloxíes de la información:** Finalmente, pa qu'una llingua tenga entrada dafechu en terrén de les tecnoloxíes del llinguaxe ye mester que la so sociedá tenga tamién accesu tecnolóxicu de mou más o menos xeneralizáu. Ésto nun significa la esistencia d'ordenadores personales en tolos llares de fala na llingua en cuestión (lo que supondría una visión occidente-céntrica dafechu del problema), sinón l'usu garantizáu de tecnoloxía informática nes instituciones académiques y d'investigación, mesmo que nes empreses privaes dedicaes al desendolcu de productos de base llingüística. Na mayoría de casos la imposibilidá d'accesu a les tecnoloxíes de la

[16] http://registry.dfki.de. Agirre *et al*. (2002) ufierta un analís d'estos datos, empobináu a les aplicaciones de PLN independientes de llinguaxe.

información vien condicionada por una seria situación de diglosia. Fai'l casu del quechua, la llingua indíxena americana más falada, con cerca de 10 millones de falantes.

Cada llingua del mundu asítiase nun puntu de coordenaes determináu en rellación a estos cuatro parámetros, lo que trai darréu un nivel determináu de capacidá de desenvolvimientu dientro de les tecnoloxíes llingüístiques. El trabayu nesta estaya puede estremase d'acordies colos tres niveles de desenvolvimientu, que correspuenden, al aldu, colos plantegaos en dellos trabayos que reflexonen sobre les estratexes pal desendolcu de tecnoloxía llingüística pa llingües minoritaries (Sarasola 2000; Agirre *et al*. 2002; Diaz de Ilarraza *et al*. 2003):

- **Fundamentos:**
  Fase d'embalagamientu de datos léxicos y corpus testuales, entá ensin procesamientu a nengún nivel. Los productos resultantes nesti estadiu tiénense de puntu de partida necesariu pal desenrueldu de tecnoloxía llingüística.

- **Ferramientes:**
  La bibliografía citada enantes considera esta fase como la empobinada a la igua de preseos pal procesamientu del llinguaxe natural: lematizadores, analizadores morfolóxicos y sintácticos, alliniadores de frases para corpus multillingües, ontoloxíes o bases de conocimientu léxico-semánticu, etiquetadores semánticos, etc. Trátase de ferramientes que tanto van sirvir nel tratamiento de la información atropada na fase previa, como van valise d'ella (por exemplu, nel entrenamientu de ferramientes de base estocástica).

  Amás de la fechura de ferramientes, al mio pensar, nesta fase hai qu'amestar tamién la ellaboración de preseos llingüísticos basaos nes colecciones d'información atropaes na etapa previa. Refiérome a diccionarios y gramátiques d'usu públicu. La so ellaboración en base a corpus textuales dexa espeyar de manera más afayadiza la llingua que se describe ende.

- **Aplicaciones:**
  Fase aplicada a la construcción d'aplicaciones empobinaes a usuarios non especializaos. Por casu: correctores gramaticales, sistemes d'ayuda a la traducción con dél nivel de sofisticadura, sistemes de busca o recuperación d'información, o sistemes de diálogu. El llabor nesta etapa mándase necesariamente de les ferramientes desendolcaes na fase anterior.

Pa les llingües con un númberu curtiu de falantes, en clara situación de diglosia y ensin sofitu institucional, la entrada na tecnoloxía de la información va ser, si non imposible, costosa de manera. Ésta ye, llamentablemente, la situación de la inmensa mayoría de les 6800 llingües qu'entá anguaño se falen nel nuesu mundu. Puestos no meyor, el so contactu cola tecnoloxía llingüística va ser pentemedies de la llingüística de corpus, si ye que daqué grupu investigador s'alcuerda d'entamar el proyecto de recoyida

de materiales léxicos ya igua de corpus (na mayoría de casos, ensin nengún tipu de procesamientu nin anotación) enantes de que la llingua en cuestión desapaeza.[17]

Nun asitiamientu intermediu afáyense, per un sitiu, les llingües pequeñes con sofitu institucional, con usu xeneralizáu a tolos niveles y una capacidá tecnolóxica prestosa, tal como holandés y finlandés. Y per otru, les grandes llingües con un grau d'accesu tecnolóxicu relativamente baxu que torgara apocayá la so entrada en campu de les tecnoloxíes llingüístiques, pero que por razones comerciales o estratéxiques tiren agora a xenerar ciertu interés. Fai'l casu del farsi o el tagalog. Nesti segundu nivel, el desenrueldu de tecnoloxía llingüística resal del nivel primariu d'atropamientu de corpus ensin etiquetar, por entrar yá na etapa de desenvolvimientu de ferramientes básiques pal procesamientu de corpus, o la igua de preseos de base llingüística (diccionarios y gramátiques) que pudieren mandase de corpus testuales básicos y bases de datos léxicos emabalagaos na etapa previa.

D'últimes, atopamos les llingües que presenten un gran númberu de falantes y que gocen de lo que calificara como un contestu de fácil accesu tecnolóxicu, como l'inglés, el chinu o l'español, daveres la ínfima minoría. Pa estos casos, el nivel de desenrueldu de les tecnoloxíes del llinguaxe ta al llegar, si entá nun lo fexo, a la creación d'aplicaciones empobinaes a usuarios non espertos como los mentaos na tercer etapa del desendolcu de la tecnoloxía llingüística.

Naturalmente, l'estremar en tres grandes bloques les llingües del mundu d'acordies col so grau de participación nes tecnoloxíes del llinguaxe ye la simplificación d'una situación enllena de matices. Un exemplu d'ello failu'l casu de les llingües minoritaries más cercanes al asturianu: el catalán, el euskera y el gallegu. Trátase de tres llingües con un númberu reducíu de falantes (unos 6 millones, 500.000 y 3,5 millones respectivamente de falantes como primer llingua) y so la influencia de daqué nivel de diglosia respectu al español (mayor o menor, según el casu), amás d'enguedeyos derivaos d'una fuerte variación dialectal nel euskera y, bien qu'en menor grau, en gallegu. Sicasí, el desenrueldu de tecnoloxía llingüística nestes tres llingües ye de sorrayase: nos tres casos se degolara la fase inicial de creación de los fundamentos y tiense entrao yá na segunda etapa, cola consiguiente xeneración de productos de base llingüística (como'l *Diccionari de freqüències* del catalán, o bien el *XX mendeko Euskararen Corpus Estadistikoa* pal euskera, consultable per internet[18]), y na construcción de ferramientes de procesamientu del llinguaxe, indispensables pa la entrada nel nivel de les aplicaciones d'usuariu non especializáu. Amás, tanto en casu del euskera como nel del catalán, comercializáronse yá delles aplicaciones finales correspondientes al tercer nivel de desendolcu (véase Diaz de Ilarraza *et al.* (2003) pal euskera, y Llisterri (2000) pal catalán). Nun dispongo d'información nesti sen no que fai al gallegu.

Cuento qu'esta situación respuende principalmente a dos factores. Per un sitiu, el compromisu cola llingua autóctona y l'enfotu de trabayar pola so normalización dende les esferes académiques y d'investigación. Otra manera, la voluntá de les administraciones nacionales catalana, vasca y gallega d'afalar el trabayu nel área, por cuantes se camienta que namá d'esta miente puede asegurase la competitividá y validez

[17] A ésto dedícase por casu *SIL International* (ente otros proyectos; véase: http://www.sil.org) o programes de mecenalgu como'l *Documentation Programme* del *The Hans Rausing Endangered Languages Project* (http://www.hrelp.org/doc_home.htm).

[18] Pa ún y otru trabayu, ver les referencies nes correspondientes secciones 3.2.2 y 3.2.4.

de la llingua nel nuevu marcu de la sociedá de la información.[19] A estos dos elementos hai que-y amestar un terceru: el nivel de desenvolvimientu de les tecnoloxíes llingüístiques de qu'esfruten estes llingües, magar la so condición, namá ye posible pa les llingües minoritaries del mundu occidental, cuidao que tienen un mínimu d'estabilidá económica garantizáu, accesu a les tecnoloxíes de la información, y tán quites d'un contestu de marxinación social como ye'l casu de les llingües indíxenes en tolos países ensin escepción del continente americanu. L'estáu de desenvolvimientu de la tecnoloxía llingüística pa les nueses tres llingües nun ye óptimu a comparalu cola situación del inglés o l' español, claramente valíes pola inversión del sector priváu. Sicasí, ye bien superior al estáu de la mayoría de llingües del mundu que s'alcuentren nuna situación comparable.

Na redolada europea, l'aplicación de les tecnoloxíes del llinguaxe nel ámbitu de les llingües minoritaries autóctones ye anguaño un tema de creciente interés. Amuesa d'ello ye la pasu ente pasu más bayurosa bibliografía sobre'l tema, la creación apocayá de SALTMIL,[20] un grupu d'interés dedicáu a ello (Nadeu *et al.* 2001), o bien los seminarios temáticos entamaos en marcu de congresos d'ampliu algame que se centraron nesta cuestión.[21] La premisa de partida ye l'aceptación de la llamada sociedá de la información como'l nuevu contestu de referencia pal truecu de conocimientu y, darréu, l'avance tecnolóxicu d'occidente, neto que'l papel indispensable de la Inxeniería Llingüística pal so progresu. Con éstes, la supervivencia de les llingües minoritaries occidentales pasa necesariamente pel desendolcu de tecnoloxía llingüística (Sarasola 2000, Nadeu *et al.* 2001).

Tiense visto que la complexón vital de cada llingua determina a grandes rasgos la so posibilidá de desenvolvese dientro de les tecnoloxíes del llinguaxe. Sicasí, tien que se considerar tamién que la entrada d'una llingua minorizada nesti campu supón, arriendes de la posibilidá d'actualización y ameyoramientu de la so presea de descripción llingüística (diccionarios y gramátiques), la so entrada en contestu de la sociedá de la información, que ye lo de calificala como llingua moderna que ye quien a competir a nivel tecnolóxicu, abandonando la so imaxe arcaica, ensin cohesión interna nin capacidá pa la comunicación a tolos niveles. En definitiva, contribui al frayamientu de la situación de diglosia na que s'afaya, lo que, otramiente, retroalimenta positivamente la voluntá de sofitu institucional y aguiya un progresivu interés comercial.[22] El desenrueldu de les

---

[19] A éstes, ye relevante la propuesta d'acción estratéxica nes áries d'Industria de la Llingua ya Inxeniería Llingüística de parte del gobiernu vascu (véase la nota 11 pala referencia ompleta), o bien la creación del *Centre de Referència en Enginyeria Lingüística* (CREL) pola Generalitat de Catalunya, y el *Centro Ramón Piñeiro para a Investigación en Humanidades* de parte de la Xunta de Galicia, dos instituciones aplicaes a la investigación y desenrueldu de tecnoloxía llingüística pa les llingües autóctones.

[20] SALTMIL ye'l *Special Interest Group on Speech and Language Technology for Minority Languages*, creau dientro de la *International Speech Communication Association* (ISCA) (http://isl.ntftex.uni-lj.si/SALTMIL/).

[21] Por exemplu los *workshops: 'Language Resources for European Minority Languages'* (Granada, LREC 1998), '*Developing language resources for minority languages: re-useability and strategic priorities'* (Atenas, LREC 2000), y '*Natural Language Processing Of Minority Languages And Small Languages'* (Batz-sur-Mer, France, TALN 2003).

[22] Nesti sen abúltame ilustrativa una anécdota tocántenes a la entrada del catalán na industria televisiva y cinematográfica: de la qu'en 1983 estrenóse TV3, la canal de televisión de Cataluña,

tecnoloxíes del llinguaxe ye, poro, daqué que va tresallá d'un cenciellu exerciciu académicu. Ye un asuntu d'interés públicu y de sonadía social, nel de xugase l'afitamientu y supervivencia de les llingües minoritaries como llingües vives nuna cultura global de base tecnolóxica.

4.2 Puntu de partida pal asturianu

La entrada del asturianu na tecnoloxía llingüística vuélvese, con éses, un proyectu de primer prioridá, oldeable a la so incorporación nos medios de comunicación de mases. La coyuntura actual paez abrir una posibilidá al proyectu: de parte de los organismos responsables del aria na Comunidá Europea, esiste una reconocencia de la situación que se-yos plantega anguaño a les llingües minoritaries autóctones nesti nuevu marcu d'información; de parte de l'administración asturiana, apúntase daqué cambiu de política tocantes a la normalización de la llingua asturiana, traducíu de recién na garantía d'aprobación d'un plan de normalización social del asturianu nun futuru próximu, y na creación de la Oficina de Política Llingüística.[23]

Amás, la llingua asturiana esfruta yá d'una mínima base de partida. De mano, dispón de l'Academia de la Llingua Asturiana, una institución con autoridá llingüística, fundamental p'articular proyectos d'investigación. De segundes, tien entrao nuna fase de normalización gracies a la publicación de la *Gramática de la Llingua Asturiana* (3ª ed., 2001) y del *Diccionario de la Llingua Asturiana* (2000) a cuenta d'esta mesma academia.Amás, dispóne de un decente cuerpu de testos escritos, sobremanera nos rexistros lliterarios y periodísticos, de los qu'una parte foi yá dixitalizada por aciu del *Proyectu Caveda y Nava*[24]*,* mesmo que dellos archivos orales, compilados nel marcu de distintos proyectos.[25] D'últimes, al tiempu d'escribir esti artículu acabose de presentar l'*Iguador de la Llingua Asturiana*, el primer correutor ortográficu pa l'asturianu, el cual ta basado nel diccionario de la academia.

Sicasí, nun tien que se faer de menos la situación enxeble na qu'anguaño s'atopa la llingua. El so estatus de non oficialidá traduzse na práctica nuna menor capacidá a la de beneficiase de sofitu económicu. Amás, esiste la constatación d'una pérda xeneracional

---

sentir falar en catalán a *JR*, *Sue Ellen* y los sos compañeros de la serie *Dallas* (yá conocíos del públicu al traviés de TVE) ye motivu de risión. En 2002 sicasí, la indignación popular y el so consiguiente boicot comercial, movíos pola refuga de la *Warner Bross* a doblar *Harry Potter* al catalán, fexo camudar la opinión de la multinacional estadounidense.

[23] Noticias publicadas en *Asturies.com, Diariu Electrónicu Asturianu* el 8 y el 29 de octubre pasado, respectivamente: http://www.asturies.com/seccion.php?fseccion=llingua#2387 y http://www.asturies.com/seccion.php?fseccion=llingua#2455 (URLs vigentes el 29 de octubre del 2003). No me consta sin embargo una mejora en la actitud del gobierno español respecto al reconocimiento de los derechos del asturiano como lengua perteneciente también al estado.

[24] http://www.cavedaynava.org/ (URL vixente el 29 d'ochobre del 2003).

[25] Por exemplu, el *Archivu Oral de la Llingua Asturiana* (http://www.asturies.org/asturianu/archoral/), el *Archivu de la Tradición Oral Asturiana*, desenroldau nel Muséu Etnográficu del Pueblu d'Asturies de Xixón y que ye la base del *Atlas sonoru de la Llingua Asturiana,* y finalmente el proyectu de compilación de testos orales llevado a cabu por parte de la Casa de la Llingua, en Corvera.

[27] http://www.asturies.org/seccion.php?fseccion=llingua#2454 (URL vixente el 29 d'ochobre del 2003).

de falantes, neto qu'un amenorgamientu del usu de la llingua mientres la última década en rexistros indagora eslusivos d'ella (Llera Ramo, 2003, según la reseña del trabajo hecha en el diario electrónico *Asturies.org*[27]). Estos dos factores suponen un acutamientu considerable en términos de recursos, y, poro, la propuesta pa un proyectu que dea entrada al asturianu nes tecnoloxíes de la información nun puede entamar per plantegamientos maximalistes. Otramiente, hai que s'atener a la esperiencia sacada de proyectos con otres llingües en situación asemeyada, como l'euskera. Nestos trabayos previos viose cómo nun ye d'empezar a desenroldar ferramientes nin aplicaciones si nun esisten los *fundamentos*; esto ye, compilaciones léxiques y corpus testuales (Sarasola 2000, ente otros).

Aplicáu al casu particular de la llingua asturiana, esto pica a la creación d'un corpus testual -escritu y/o oral- como prioridá inicial pa la so entrada en campu de les tecnoloxíes llingüístiques. Un corpus testual, en xunto col atropamientu de datos léxicos (que puede tener el diccionariu de la ALlA como puntu de partida), garantiza l'encontu pal desenvolvimientu posterior de ferramientes y aplicaciones. De la mesma, permite a mediu plazu la fechura de productos de base llingüística. Por casu, los estudios llingüísticos realizaos sobre los datos del corpus pueden derivar n'aplicaciones como l'ameyoramientu de la gramática y el diccionariu mentaos enantes, la creación d'un diccionariu descriptivu a cuenta del léxicu más frecuente de la llingua, o bien la planificación de delles actuaciones sociollingüístiques. Otramiente, ye posible pensar n'aplicaciones que valgan al aprendizaxe del asturianu: cuando ameyorando los materiales esistentes a partir de datos procedentes del corpus, cuando desenroldando ferramientes pal aprendizaxe valíu por ordenador –por exemplu, exercicios que requieran reproducir les categoríes morfosintáctiques asignaes en determinaes frases del corpus, xenerar formes flexonaes a partir d'un lema y la información a él venceyada, completar perífrasis, frases feches y allugamientos, etc.

Nenguna d'estes aplicaciones quier un nivel de tecnoloxía sofisticáu por demás. Un corpus testual lematizáu y etiquetáu morfosintácticamente, en xunto a una ferramienta d'esplotación de los datos, fai abondo. Asitiámonos namá nel segundu nivel de desenrueldu tecnolóxicu presentáu na sección 4.1, y, sicasí, les posibilidaes de les aplicaciones que se planteguen son yá notables.

Repárese, con eso, que se trata d'aplicaciones que presuponen un fragmentu particular de lo que ye la llingua asturiana: l'asturianu actual. La entrada del asturianu nes tecnoloxíes del llinguaxe plantégase (arriendes de pol inherente interés a nivel científicu) en fieces a derrompe-y la entrada al nuevu marcu de comunicación y llexitimalu en cuantes que llingua válida en tolos niveles d'usu. Teniendo de cuenta la so presente situación, un corpus de llingua actual abulta un beneficiu inicial más xeneralizable a distintes aries de la llingüística aplicada que non, por casu, un corpus históricu.

Otra manera, ye relevante de manera qu'esti primer corpus potencial recueya la máxima variedá dialectal, diastrática y diafásica del asturianu, por mor del estadiu d'asimilación de la norma estándar nel que s'alcuentra anguaño, a pocos años de l'apaición d'un diccionariu y una gramática asoleyaos pola institución con autoridá normativa sobre la llingua. Cuandoquier, interesa qu'esti primer corpus de la llingua asturiana sía un corpus de referencia de llingua xeneral.

Un corpus pal asturianu tien que se concebise tamién con cuenta de que, a la llarga, se mande como base pal desenrueldu d'aplicaciones tecnolóxiques más sofisticaes: sistemes

de diálogu, sistemes de sumarización, de recuperación o cata d'información, etc. D'ésta, son mester dos elementos. D'uno, que les ferramientes básiques de procesamientu, como lematizadores, analizadores morfosintácticos o sintácticos, se fixeren de mou que puedan ser bonos de reutilizar, a faese necesario, de parte de caúna d'estes aplicaciones finales (nuevamente Sarasola 2000 y Diaz de Ilarraza *et al.* 2003 defenden esti criterio, resultanza de la so esperiencia col euskera). De segundes, que la igua d'estes aplicaciones finales surda d'un analís de la realidá del asturianu y la so adecuación a ella.

Esto ye: chando cuenta de la condición de llingua minoritaria del asturianu, nun ye dable plantegar el desendolcu de toles aplicaciones nes que se trabaya anguaño pa llingües como l'inglés. Pela cueta, hai que concentrar l'esfuerzu naquelles xeres pa les qu'esista garantía real d'usu –y, con ella, un mínimu de rentabilidá comercial. Considerando les condiciones de la llingua asturiana, abulta claro que, por exemplu, un sistema de sumarización de documentos va a ser d'un interés y nivel d'aplicación enforma inferior al que puede tener un sistema de traducción automática ente l'asturianu y l'español. Poro, el proyectu de construcción d'un corpus de llingua asturiana tien de considerar, na mio opinión, la posibilidá de que siquier una parte sea billingüe asturiano-español, preferentemente paralelo.

Finalmente, el proyectu tamién tien de char cuenta del estáu nel que s'afaya el fragmentu de llingua que mire a representase. Por exemplu, va tener que tomar decisiones sobro'l tratamientu de la variación dialectal; por exemplu, la qu'afecta a la flexón de determinaes categoríes, pola repercusión qu'esto puede tener en procesu de lematización y asignación de etiquetes morfosintáctiques. O, otra manera, va tener qu'establecer una política específica pa la variación ortográfica, a la de tratar con testos d'una etapa pre-normativa. Nesti sen, puede sirvir de referencia la esperiencia d'otros proyectos desendolcaos pa llingües cercanes al asturianu, como les citaes.

5 Conclusiones

Nesti artículu plantegóse la posibilidá d'un proyectu de corpus pa la llingua asturiana. Con esti envís, presenté brevemente l'oxetu d'estudiu de la Llingüística de Corpus y l'interés qu'ufierta dientro de les disciplines de base llingüística y la investigación como pa la igua de ferramientes y aplicaciones. Viose que l'enclín fonderu d'un corpus llingüísticu ye sirvir de fonte de datos pal analís empíricu d'una llingua y la so consiguiente modelización, que puede ser utilizada en distintos niveles de desendolcu: n'estudios de base teórica, na ellaboración de presea de base llingüística como diccionarios o gramátiques descriptives, o pa la fechura de ferramientes de base tecnolóxica, como correctores ortográficos y gramaticales o programes de traducción automática.

Un corpus llingüísticu ye, poro, d'induldable interés tanto pa la comunidá de profesionales del llinguaxe (dende llingüistes teóricos y sociollingüistes, hasta lexicógrafos y traductores), neto que, a más llargu plazu, pal público xeneral, beneficiariu de les aplicaciones resultantes. Toa esta potencialidá fai de los corpus llingüísticos la basa necesaria pal desendolcu de tecnoloxía llingüística en cualaquier llingua.

Otramiente, tien analizaose la rellación ente llinguas minorizaes y tecnoloxíes del llinguaxe, y viemos como en marcu actual de la sociedá de la información, les tecnoloxíes llingüístiques van xugar un papel definitivu a la d'algamar una normalización

social dafechu de les llingües minorizaes. Obviamente, esto ye posible en cuantes que se satisfazan delles condiciones, como un accesu tecnolóxicu mínimamente garantizáu o'l sofitu institucional a la llingua.

Magar la so enxeble situación, el so marcu social y institucional anguaño paez cumplir mínimamente talos requisitos. Ye importante entós aprovechar la coyuntura y asegurar la entrada de la llingua nes tecnoloxíes del llinguaxe, que tien de pasar necesariamente pela construcción d'un corpus del asturianu. Desque asumío esto, yá se suxurieren delles característiques que debería cumplir esti corpus potencial col envís de garantizar un rendimientu másimu magar la limitación de recursos na que s'afaya la llingua. Específicamente, consideróse qu'un primer corpus pal asturianu tien que se concebir como un corpus de llingua xeneral, que dea cuenta de les sos variedaes diatópiques, diastrátiques y diafásiques, y que, a poder ser, incluya tamién un fragmentu de testos paralelos asturianu-español.

6 Bibliografía

ABAITUA, JOSEBA (1999) "Quince años de traducción automática en España". *Perspectives: Studies in Translatology*. 7-2: 221-230. Versió catalana: "Quinze anys de traducció automàtica a l'Estat espanyol". *Digit.HUM*. Universitat Oberta de Catalunya. http://www.uoc.es/humfil/digithum/digithum2/catala/didactica/index.html

AGIRRE E., ALDEZABAL I., ALEGRIA I., ANSA O., ARREGI X., ARRIOLA J., ARTOLA X., DÍAZ DE ILARRAZA A., EZEIZA N., GOJENOLA K., MARITXALAR A., MARITXALAR M., ORONOZ M., SARASOLA K., SOROA A., URIZAR R., URKIA M. (1998) 'A framework for the automatic processing of Basque'. *Workshop on Lexical Resources for Minority Languages*. LREC 1998, Granada.

AGIRRE E., ALDEZABAL I., ALEGRIA I., ARREGI X., ARRIOLA J.M., ARTOLA X., DÍAZ DE ILARRAZA A., EZEIZA N., GOJENOLA K., SARASOLA K., SOROA A. (2002) 'Towards the definition of a basic toolkit for HLT'. *LREC 2002. Third International Conference On Language Resources And Evaluation.* Las Palmas, Islas Canarias.

AGUIRRE MORENO, J.L., N. ANDIÓN, X. GÓMEZ GUINOVART (2001) "Aspectos ortográficos, léxicos y morfosintácticos del etiquetado lingüístico de un corpus de informática en lengua gallega". *Procesamiento del Lenguaje Natural*, 27: 13-19.

AGUIRRE MORENO, J.L., A. ÁLVAREZ LUGRÍS, X. GÓMEZ GUINOVART (2002) "Etiquetario morfosintáctico del SLI para corpus de lengua gallega: aplicación al corpus paralelo TECTRA". *Procesamiento del Lenguaje Natural*, 28: 23-34.

AGUIRRE MORENO, J.L., A. ÁLVAREZ LUGRÍS, I. BRAGADO TRIGO, L. CASTRO PENA, X. GÓMEZ GUINOVART, A. LÓPEZ LÓPEZ, J. R. PICHEL CAMPOS, E. SACAU FONTENLA, L. SANTOS SUÁREZ (2003) "Alinhamento e etiquetagem de corpora paralelos no CLUVI (Corpus Linguístico da Universidade de Vigo)". Almeida, J.J. (ed.), *Workshop CP3A 2003, Corpora Paralelos: Aplicações e Algoritmos Associados*. Universidade do Minho, Braga, Portugal.

ASTON, GUY, LOU BURNARD (1998) *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh University Press.

BACH, CARME, ROSER SAURÍ, JORDI VIVALDI, M. TERESA CABRÉ (1997) *El Corpus de l'IULA: descripció*. Sèrie Informes. IULA, Barcelona.

BIBER, D. (1993). 'Representativeness in corpus design'. *Literary and Linguistic Computing*, Vol. 8 (4): 243-257.

BOIX, E. (1996) 'Els materials de llengua oral dels corpus de català contemporani de la UB (CUB)'. LL. PAYRATÓ, E. BOIX, M.R. LLORET, M. LORENTE (Eds.) *Corpus, Corpora. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2)*. Barcelona: Promociones y Publicaciones Universitarias SA. pp. 93-114.

CARLSON, LYNN, DANIEL MARCU, MARY ELLEN OKUROWSKI (2003) 'Building a Discourse-Taggeed Corpus in the Framework of Rethorical Structure Theory'. KUPPEVELT, JAN VAN, RONNIE SMITH (Eds.) *Current Directions in Discourse and Dialogue*. Kluwer (to appear).

DIAZ DE ILARRAZA A., A. GURRUTXAGA, I. HERNAEZ, N. LOPEZ DE GEREÑU, K. SARASOLA (2003) 'HIZKING21: Integrating language engineering resources and tools into systems with linguistic capabilities'. *Workshop on NLP of Minority Languages and Small Languages.* TALN 2003. Batz-sur-Mer, 11-14 juin 2003.

FLIGELSTONE, STEVE (1992) 'Developing a scheme for annotating text to show anaphoric relations'. Gerhard Leitner (Ed.) *New Directions in English Language Corpora: Methodology, Results, Software Developments*. Berlin: Mouton de Gruyter.

FRANCIS, W.N., H. KUCERA (1964) *Manual of Information to Accompany a Standard Sample of Present-day Edited American English, for use with Digital Computers*. Providence RI: Department of Linguistics, Brown University. Edición revisada y ampliada (1979): http://www.hit.uib.no/icame/brown/bcm.html (última actualización: 11/9/1997).

GARCÍA-MATEO, CARMEN, MANUEL GONZÁLEZ-GONZÁLEZ (1998) 'An overview of the existing language resources for Galician'. *Workshop on Language Resources for European Minority Languages*. LREC, Granada, 28-30 May 1998

LLERA RAMO, FRANCISCO JOSÉ (2003) 'Lluces y solombres na evolución sociollingüística del asturianu'. *XXII Xornaes Internacionales d'Estudiu de la llingua Asturiana*. Academia de la Llingua Asturiana, Vigo, 27-29 ochobre 2003.

LLISTERRI, JOAQUIM (1997) 'Transcripción, etiquetado y codificación de corpus orales'. *Seminario de Industrias de la Lengua,* Fundación Duques de Soria, 15 de julio de 1997.

LLISTERRI, JOAQUIM (1999) "Corpus orals per a la fonètica i les tecnologies de la parla". *Actes del I Congrés de Fonètica Experimental*. Tarragona, 22, 23 i 24 de febrer de 1999. Universitat Rovira i Virgili - Universitat de Barcelona: 27-38.

LLISTERRI, JOAQUIM (2000) 'O catalán nas industrias da lingua', *Lingua e cultura catalanas*, Cursos de extensión universitaria, Universidade de Vigo, 5 July 2000.

MCENERY, TONY, ANDREW WILSON (2001) *Corpus Linguistics*. Edinburgh University Press.

MANN, WILLIAM, SANDRA THOMPSON (1988) 'Rethorical Structure Theory. Toward a functional theory of text organization'. *Text*, 8(3): 243-281.

NADEU, CLIMENT, DONNCHA Ó'CRÓINÍN, BOJAN PETEK, KEPA SARASOLA, BRIONY WILLIAMS (2001) ISCA SALTMIL SIG: Speech and Language Technology for Minority Languages. http://gps-tsc.upc.es/veu/research/pubs/download/Nad01b.pdf (28/10/2003)

PINO, M., M. SÁNCHEZ SÁNCHEZ (1999) "El subcorpus del Banco de Datos CREA-CORDE (RAE): procedimientos de transcripción y codificación". *Oralia*, 2: 83-138.

PUSTEJOVKSY, JAMES, PATRICK HANKS, ROSER SAURÍ, ANDREW SEE, ROBERT GAIZAUSKAS, ANDREA SETZER, BETH SUNDHEIM, LISA FERRO (2003) 'The TIMEBANK Corpus'. *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster, 28-31 March 2003.

QUIRK, R. (1992). "On Corpus Principles and Design". JAN SVARTVIK (ed.) (1990) *The London Corpus of Spoken English: Description and Research.* Lund University Press: 457- 469.

QUIRK, R., S. GREENBAUM, G. LEECH AND J. SVARTVIK (1985) *A comprehensive Grammar of the English Language.* Harlow, Longman.

RAFEL, J. (1992-93) "El 'Diccionari del català contemporani': Treballs realitzats i previsions de futur", *Llengua i Literatura* 5: 733-737.

RAFEL, J. (1996) "El Diccionari del català contemporani i el Corpus textual informatitzat de la llengua catalana". LL. PAYRATÓ, E. BOIX, M.R. LLORET, M. LORENTE (Eds.) *Corpus, Corpora. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2)*. Barcelona: Promociones y Publicaciones Universitarias SA: 71-92.

RAFEL I FONTANALS, J. (Dir.) (1996-1998) *Diccionari de freqüències*. 3 Vols. Barcelona: Institut d'Estudis Catalans. CD-ROM de l'obra completa.

RODRÍGUEZ, CARLOS (2003) "Applying Information Extraction techniques to metalinguistic discourse", *Topics in Computational Linguistics and Intelligent Text Processing*; Lecture Notes in Computer Science. Springer-Verlag. (en prensa)

SÁNCHEZ, A., R. SARMIENTO, P. CANTOS, J. SIMÓN, J. (1995) *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones*. Madrid: SGEL.

SÁNCHEZ-LEÓN, F., J. PORTA, J.L. SANCHO, A. NIETO, A. BALLESTER, A. FERNÁNDEZ, J. GÓMEZ, L. GÓMEZ, E. RAIGAL, R. RUIZ (1999). «La anotación de los corpus CREA y CORDE». *Actas del XV Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, SEPLN.

SARASOLA, K. (2000) 'Strategic priorities for the development of language technology in minority languages'. *Workshop on Developing Language Resources for Minority Languages: Re-useability and Strategic Priorities*. LREC 2000. Athens, Greece.

SINCLAIR, J. (ed.) (1987) *Looking Up, An Account of the COBUILD Project*. London: Collins.

SINCLAIR, J. (1996) *Preliminary recommendations on corpus typology*. EAGLES Document TCWG-CTYP/P. http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html (21/9/03)

SOLER I BOU, J. (1998) 'Los corpus textuales en lengua catalana'. *Curso de Industrias de la Lengua "Proyectos actuales en procesamiento de lenguaje natural"*, Fundación Duques de Soria, Soria, 13-17 de julio de 1998.

SOMERS, HAROLD (2001) 'Where do we stand?'. Panel Session, *MT Summit VIII*, Santiago de Compostela. 18-22 September 2001.

SVARTKIK, JAN (ed.) (1990) *The London Corpus of Spoken English: Description and Research*. Lund Studies in English 82. Lund University Press.

TOGNINI-BONELLI, E. (1996). *Corpus Theory and Practice*. Birmingham: TWC.

TORRUELLA, JOAN, JOAQUIM LLISTERRI (1999) 'Diseño de corpus textuales y orales'. BLECUA, J.M., G. CLAVERIA, C. SÁNCHEZ, J. TORRUELLA (Eds.) *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona, Universitat Autònoma de Barcelona, Editorial Milenio: 45-77.

VIAPLANA, J. (2000) 'Corpus oral de variació'. *Jornades del Centre de Referència en Enginyeria Lingüística (CREL)*, Institut d'Estudis Catalans, Barcelona, 4 i 5 d'abril de 2000.