Sondeo neológico en la prensa digital en lengua asturiana (Asturnews 2004-2008)

Proyecto final de máster

Màster en Lingüística i Aplicacions Tecnològiques
(Universitat Pompeu Fabra, Barcelona)

Índice

Prólogo

El 10 de julio pasado, el PSOE y el PP —fuerzas dominantes en la Xunta Xeneral del Principáu d'Asturies, con 41 de sus 45 diputados— anunciaban un acuerdo por el que dejaban cerradas las líneas maestras de lo que habrá de ser el nuevo Estatuto de Autonomía de Asturies[i].

Entre otros asuntos, el pacto alcanzado zanjaba definitivamente —al menos en intención— una de las cuestiones que han ocupado más espacio en el debate político de Asturies en las últimas décadas. Me refiero a la cooficialidad del asturiano y la fala eonaviega *con el castellano, lengua mayoritaria del Principado.*

La forma de dar carpetazo a la cuestión no pudo ser más decepcionante para la causa asturianista. *La prerredacción del texto estatutario no sólo dejaba fuera de juego toda opción de oficialidad sino que venía a apuntillar los pocos avances conseguidos en treinta años de reivindicación legal, política y social. Hay quien opina, incluso, que lo pactado supone un retorno a principios legislativos tardofranquistas*[ii].

Sea como fuere, sí está claro que, con su propuesta, los grandes partidos asturianos han creado una situación de anomalía total: respecto a las directrices de sus direcciones nacionales, favorables, al menos formalmente, al plurilingüismo del Estado; respecto a los marcos legales de otras comunidades autónomas con lengua propia[iii] *y respecto a distintos tratados internacionales suscritos a bombo y platillo por España ante organismos como el Consejo de Europa o la UNESCO.*

El principio de acuerdo, entre otras cosas, consolida la condición de patrimonio *y* valor de convivencia *de las lenguas autóctonas de Asturies, pasa por alto la legislación existente (una* Llei d'usu *firmada en 1998 pero nunca puesta en vigor) e impone unas condiciones draconianas para la elaboración de una nueva ley de uso: el voto a favor de dos terceras partes del parlamento asturiano.*

Estaríamos, en suma, ante toda una proclama de la hegemonía política asturiana. Por si a alguien le quedaba alguna duda, la Administración autonómica se ha instalado definitivamente en el terreno de la contraplanificación[iv]*, esto es, en el de las proclamas vacuas que, en la práctica, esconden formas de auténtica represión.*

Ahora toca, por fin, aceptar el hecho de que los legisladores del Principáu no son la solución a la lenta y vergonzosa agonía de la llingua*, sino más bien una de sus causas directas.*

Ante esta situación de abandono, con un poder político a la contra y unas instituciones oficiales muy mermadas en sus capacidades[v]*, los hablantes han terminado asumiendo, sin siquiera percatarse de ello, una responsabilidad involuntaria: la de planificadores lingüísticos en precario. Son ellos quienes, con sus*

esfuerzos diarios, deben mantener —y, de hecho, mantienen— en movimiento la rueda de la normalidad lingüística[vi].

Ahora bien, como pretendo demostrar indirectamente con el trabajo al que antecede esta suerte de introducción vehemente*, Internet —y, de forma muy especial, su subproducto, la red social o Web 2.0— parece estar haciendo de esa precariedad un arma para conseguir lo que la vía política sigue negando: unas condiciones que permitan la pervivencia del asturiano como lengua* normal *en el uso social.*

En efecto, el desarrollo vivido por Internet y la Web 2.0, de la que el diario Asturnews*, con su carácter participativo y su espíritu* copyleft*, es uno de sus más marcados exponentes en lengua asturiana, parece ser uno de los motivos por los que, pese a haber retrocedido socialmente, el asturiano habría ganado en prestigio y en presencia en la Red.*

Los ciberfalantes *no son muchos, pero sí hacen mucho ruido. Así, aunque Asturies es una comunidad muy envejecida, con la menor proporción de población infantil de España (10,1 % de habitantes menores de 15 años) y la segunda mayor de población anciana (22,1 % mayores de 64 años), ocupa el sexto lugar en cuanto a penetración de banda ancha, con 17,8 líneas por cada 100 habitantes (por encima de la media nacional, de 17,7 líneas)*[vii].

En este contexto, logran entenderse mejor datos como las casi 300.000 visitas que, durante los últimos meses, habría recibido de media Asturnews*, sobre un público potencial de 240.000 lectores*[viii] *o el hecho de que la Wikipedia asturiana (*Uiquipedia*), con unas 12.000 entradas, juegue en la segunda división de las enciclopedias libres, algo extraordinario para una* lengua de tercera *en cuanto a derechos como el asturiano.*

Internet y la Web 2.0 son —no me cabe la menor duda— una de las piezas de las que depende el futuro del asturiano. Los hablantes parecen estar dándose cuenta de ello, como queda patente por el ritmo con que la llingua *se va incorporando a las innovaciones que ofrece la Red: prensa digital, foros de discusión, buscadores, tiendas* online*, listas de distribución, entradas de la enciclopedia libre, blogosferas, proyectos de desarrollo de software libre, marcadores sociales, canales de vídeo en YouTube, campañas virales de reivindicación lingüística*[ix]*…*

Bastaría, pues, un poco de organización, una toma de conciencia del potencial planificador que ofrece la Red, para empezar a conseguir efectos inmediatos de gran relevancia, capaces de reactivar la rueda de la normalización.

Lejos de dejarnos llevar por el entusiasmo, no obstante, debe primar la prudencia, ya que existe un riesgo serio de terminar creando esferas paralelas, ajenas a la realidad social de ahí fuera*. Dicho en otras palabras: hay que evitar a toda costa que se produzca un cisma entre el* asturiano online*, escrito y virtual, y el asturiano de la calle, oral y mestizo, ya que, entonces, estaríamos jugando con fuego.*

La mejor forma de controlarlo, en mi opinión, es buscar algún referente mesurable, que aporte objetividad. Y el mejor candidato para ello es la neología, las palabras nuevas que, día a día, van apareciendo, van señalando necesidades y, ante todo, nos van indicando la vitalidad que le queda a la maltrecha lengua asturiana.

Es necesario, por tanto, hacer un inventario neológico representativo, fiable y actual; y es necesario hacerlo cuanto antes. Ése es el compromiso que ha impulsado el trabajo académico que viene a continuación. Espero tan sólo que el esfuerzo termine mereciendo la pena.

Alejandro Fernández-Cernuda Díaz
Septiembre del 2008, Barcelona

[i] Por su repercusión, los medios asturianos recogieron profusamente la noticia; fuera de Asturies, como suele suceder con todo lo relativo a la *llingua*, simplemente pasó desapercibida. Se indican a continuación las URL sobre la noticia de los principales periódicos asturianos (*La Nueva España* y *La Voz de Asturias*):
http://www.lne.es/secciones/noticia.jsp?pRef=2008071100_42_655447_
Asturias-capitalidad-ultimo-escollo-para-PSOE-sellen-pacto-sobre-Estatuto
http://www.lavozdeasturias.es/noticias/noticia.asp?pkid=335086

[ii] De acuerdo con un informe publicado el 14 de julio del 2008 por el Aconceyamientu de Xuristes pol Asturianu (AXA), la propuesta firmada por el PSOE y el PP tendría un contenido y un espíritu calcados del Decreto 2929/1975, aprobado en los estertores de la dictadura franquista: «Las lenguas regionales son patrimonio cultural de la Nación española. […] Su conocimiento y uso será amparado y protegido por la acción del Estado».

[iii] De las comunidades autónomas con lenguas propias en todo su territorio (Galicia, Catalunya, Illes Balears y Asturies), sólo Asturies ha renunciado en su estatuto a la declaración de cooficialidad, con lo que queda equiparada a Aragón o Castilla y León, monolingües y castellanohablantes en la mayoría de su extensión.

[iv] Tomo este término en el sentido en que el Dr. Brauli Montoya Abat (Universitat de Alacant) lo aplica al caso valenciano en su obra *Normalització i estandardització* (2007, Edicions Bromera).

[v] Me refiero, por un lado, a la escasez de recursos y de autoridad que sufren la Academia de la Llingua Asturiana, la Oficina de Política Llingüística o la Rede de Conceyos pola Normalización y, por otro, a la falta de una línea clara de actuación respecto a la *llingua* que caracteriza a la Universidá d'Uviéu.
Un ejemplo extremo de la *tibieza* con que la Universidá d'Uviéu trata la cuestión del asturiano —pese al compromiso expreso que recogen sus estatutos— se dio este verano (24 de junio) con la exclusión, en una votación muy cuestionable, del asturiano como asignatura para los futuros títulos de grado. El saldo de la polémica fue la dimisión inmediata de la decana de la Facultá de Filoloxía, la Dra. Ana Mª Cano González (también presidenta de la Academia de la Llingua Asturiana) y la confirmación de la honda división ideológica que existe en la comunidad académica asturiana. Para más información sobre este incidente:
http://www.lne.es/secciones/noticia.jsp?pRef=2008062500_46_650439__SOCIEDAD-Y-CULTURA-
Filologia-Oviedo-suprime-Asturiano-decana-Cano-dimite

[vi] Aquí me inspiro en el modelo de *Catherine Wheel*, desarrollado por el Dr. Miquel Strubell i Trueta (Universitat Oberta de Catalunya) y con el que se representa, con un ciclo de fases interrelacionadas y retroalimentadas, el ideal de todo proceso de planificación lingüística: a más aprendizaje en la lengua minoritaria, más demanda de bienes y servicios en ella, más oferta de productos, más consumo, mayor percepción de utilidad en su uso, mayor motivación para aprender y utilizarla, más aprendizaje…, y así sucesivamente.

[vii] Fuentes: Instituto Nacional de Estadística (INE) y Comisión del Mercado de Telecomunicaciones (CMT).

[viii] El 22 % de habitantes del Principáu que afirma entender, hablar y leer el asturiano, de acuerdo con el *II Estudio sociolingüístico de Asturias (2002)* del Dr. Francisco Llera Ramos (Euskal Herriko Unibertsitatea).

[ix] Cito, por orden, algunos ejemplos: *Asturies.com*, *Asturnews* y *Les Noticies* (prensa digital y foros de discusión); *Úlos* (buscador); *Astur*shop (tienda *online*); *Llingua-list* y *Léxicu internacional* (listas de distribución); *Uiquipedia*; canal *Blogues* de *Asturies.com*, *Altuxa blogs* (y wiki) y *Nireblog n'asturianu* (blogosferas); *Softastur* y *Ubuntu Asturian Translators* (software libre); *Esbilla.net* (marcador social); canal *terapiadegrupoTPA* en YouTube, con casi 600 vídeos; campaña *Doilacara.net* del Conceyu Abiertu pola Oficialidá (CAO)…

0. Resumen

Pese a su delicada situación, motivada por la ausencia de un marco legal adecuado, la lengua asturiana vive una época de relativo esplendor en varios terrenos, vinculados principalmente a la esfera cultural. Así, frente al innegable retroceso social, la *llingua* ha ido ganando en los últimos años en visibilidad y prestigio. Internet, sin duda, ha desempeñado un papel fundamental en este proceso, sobre todo a raíz del desarrollo del concepto de red social o Web 2.0.

Ahora bien, cabe plantearse la cuestión: ¿esta *medionormalidad* es real o impostada? Si analizamos los datos sociolingüísticos de alfabetización y consumo de productos en asturiano, se hace difícil disipar el riesgo de que se acabe derivando hacia una endogamia artificial, donde unos pocos terminen escribiendo y creando para sí mismos, ajenos los usos lingüísticos de la sociedad.

Es aquí precisamente donde entra en juego la neología, como índice tanto de la vitalidad interna de la lengua —aspecto clave en situaciones de minorización y retroceso en el uso—, como del grado de extrañamiento que necesariamente acompaña a todo proceso de estandarización y normalización.

Mi trabajo, por tanto, se plantea desde un enfoque neológico. Así, mediante un sondeo realizado a partir de un corpus representativo extraído de uno de los medios digitales con mayor prestigio y difusión en asturiano, *Asturnews*, intentaré, por un lado, reflejar la capacidad creativa de la *llingua* y, por otro, favorecer la reflexión teórica y práctica en un terreno que considero esencial para su evolución y supervivencia. Una vez completo, además, el corpus habría de integrarse en el proyecto Eslema (corpus electrónico general de la lengua asturiana) de la Universidá d'Uviéu, con la meta última de que sirva para enriquecer con nuevos recursos digitales la tarea de otros investigadores.

La metodología de trabajo, tomada del modelo del OBNEO-IULA (Observatori de Neologia, Institut de Lingüística Aplicada) de la Universitat Pompeu Fabra —referente dentro y fuera del ámbito románico—, implica la utilización de un corpus lexicográfico de exclusión y de estrategias de ingeniería lingüística para el vaciado y la exclusión de los neologismos. Para ello, recurriré al texto completo del diccionario normativo de la Academia de la Llingua Asturiana (ALLA), base de la neología moderna en lengua asturiana, y a un conjunto de herramientas informáticas habituales en el sector y diseñadas *ad hoc*.

El resultado será el punto de arranque de mi posterior investigación doctoral, con la que intentaré pasar del sondeo a la auténtica *prospección* neológica. Para ello, abordaré una adaptación de las herramientas y estrategias de extracción y lematización del IULA, desarrollaré otras más específicas para mi caso de estudio y enriqueceré el corpus de extracción con nuevas fuentes.

Palabras clave
asturiano, lenguas minoritarias, normalización, neología, corpus electrónico, ingeniería lingüística

1. Puntos de partida de la investigación

1.1 Definición del problema

El *problema* o núcleo central de mi proyecto surge de la detección de tres nichos de investigación convergentes en torno a la lengua asturiana.

En primer lugar, estaría la ausencia de estudios exhaustivos sobre neología en asturiano basados en datos empíricos sobre el uso real. En efecto, aunque la neología asturiana ha absorbido no pocos esfuerzos en los últimos años, el enfoque dado hasta la fecha ha sido preeminentemente prescriptivo y planificador, en forma de propuestas para adecuar la *llingua* a las exigencias de los tiempos.

Por otro lado, tendríamos el retraso con que se está incorporando el asturiano a los avances en neología. Este retraso sería común a otras esferas de investigación lingüística y estaría motivado por la situación de anomalía en que se encuentra la *llingua*, que exige a los especialistas una dedicación extra a temas tan cruciales como la supervivencia o el fomento de usos lingüísticos.

Finalmente, mi propuesta buscaría cubrir, al menos de forma parcial, el hueco en materia de recursos de ingeniería lingüística que, en su condición de lengua minoritaria[1], sufre el asturiano. Con ello buscaría sumarme a otros trabajos que se han venido desarrollando en los últimos años. La finalidad aquí sería doble: trasladar al asturiano algunas de las herramientas y aplicaciones que han probado su eficacia en lenguas próximas, como el catalán o el castellano, y realizar una aportación al proyecto Eslema, con el que el Seminariu de Filoloxía de la Universidá d'Uviéu, bajo la experta dirección del Dr. Xulio Viejo Fernández, está componiendo el primer corpus electrónico general de la *llingua*[2].

Así pues, mi proyecto de corpus, elaborado con datos actuales (2004-2008) y de uso real extraídos del diario digital *Asturnews*, busca, por un lado, dar respuesta al primer nicho, respuesta que se vería reforzada por el empleo de un corpus lexicográfico de extracción basado en el texto completo del *Diccionariu de la llingua asturiana* (DALLA), de la Academia de la Llingua Asturiana (ALLA). La relevancia de esta obra está clara: su publicación, en el año 2000, no sólo culminó el proceso normativizador arrancado por la ALLA en su fundación, en 1981[3], sino que, con sus casi 52.000 entradas lexicográficas, también supuso la mayor sistematización y renovación léxica asumida hasta la fecha en asturiano.

La aplicación, una vez extraídos los datos de trabajo, del modelo metodológico del Observatori de Neologia (OBNEO) y de algunas de las estrategias

[1] Intento recoger aquí la idea que refleja el término inglés *less-resourced languages* —lenguas pobres de recursos—, semejante, en el ámbito de la lingüística computacional y la ingeniería lingüística, a los conceptos sociolingüísticos de *lengua minoritaria* y *lengua minoritaria*. En realidad, al caso del asturiano le vendrían bien los tres calificativos de *minoritaria*, *minorizada* y *pobre de recursos*.

[2] Para más información sobre este proyecto: http://www.uniovi.es/eslema.

[3] La Ortografía (1981) y la Gramática (1998) oficiales, que ya han vivido varias revisiones y reediciones, serían los otros grandes hitos de este proceso.

desarrolladas para éste por el Laboratori de Tecnologies Lingüístiques (LATEL), ambos grupos vinculados al reconocido Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra, pretende, por otro lado, poner al asturiano en condiciones de comparar su neología con la que se recoge actualmente en lenguas hermanas como el castellano, el catalán o el gallego y sentar las bases para todo tipo de estudios contrastivos en algunas de las líneas de mayor interés en la neología actual.

Por último, la adaptación —parcial en la fase del sondeo— de herramientas y estrategias tecnológicas de extracción y detección de neologismos, la propia confección del corpus de Asturnews y la explotación de recursos ya existentes, como la edición digital del DALLA y el Iguador (corrector ortográfico gratuito de la ALLA)[4], pretende servir de apoyo a los esfuerzos que, desde diversos focos, se están llevando a cabo para dotar al asturiano de unos recursos mínimos de explotación computacional.

[4] Ambos trabajos, muestra del compromiso de la ALLA con la creación de recursos digitales, son fruto de la estrecha colaboración que existe entre la institución académica y Araz, primera empresa de servicios tecnológicos en lengua asturiana y creadora, entre otros, del primer periódico digital, el primer buscador y la primera tienda *online* en asturiano (Asturies.com, Úlos y Astursho p, respectivamente).
El Iguador y la versión digital del DALLA están disponibles en http://www.academiadelallingua.com.
Para más información sobre Araz: http://www.araz.net.

1.2 Antecedentes de estudio

La *novedad* de mi trabajo será, precisamente, la búsqueda de la convergencia entre varios esfuerzos previos en torno a la neología en asturiano, de ahí que la referencia a los antecedentes deba ser prioritaria.

En este apartado comentaré los antecedentes que sitúan en contexto mi sondeo, tanto en el entorno de la *llingua* asturiana, como en los de la neología y la ingeniería lingüística.

Llingua asturiana

Fuera de los logros de la neología popular y espontánea, que, por fortuna, sigue muy viva, de los hallazgos de la literatura y de obras y estudios puntuales realizados en los ochenta, muchos de ellos de carácter institucional[5], es un hecho aceptado que la publicación del DALLA en el año 2000 supuso el punto de partida de la neología moderna en lengua asturiana.

Como obra normativa por excelencia, la mayor completada hasta la fecha en asturiano[6], el DALLA sirvió para consagrar parte de la neología culta y popular aparecida durante el Surdimientu[7] (D'Andrés: 2008) y para descartar otras innovaciones que tenían escaso fundamento. Sin embargo, su mayor aportación al terreno de la neología fue la del *criterio*: dictó los principios que guían hoy día la viabilidad —e inviabilidad— de las nuevas palabras asturianas.

Con la misma intención prescriptiva y normalizadora del DALLA, la ALLA arrancó en el año 2005 una colección de documentos neológicos (*Cartafueyos normativos*; Academia de la Llingua Asturiana: 2005a, 2006a, 2006b y 2007) con los que pretendía satisfacer las nuevas necesidades denominativas —del área de la administración sobre todo— aparecidas a raíz de la constitución de la Oficina de Política Llingüística del Principáu d'Asturies[8] y de la Rede de Conceyos pola Normalización (red de concejos con oficinas de política lingüística), en los años 2003 y 2004, respectivamente.

En esa misma línea, tanto la Oficina de Política Llingüística como las oficinas municipales —entre las que destaca, por su importancia, la de Xixón, principal núcleo urbano de Asturies, con unos 275.000 habitantes[9]— desempeñan una notable tarea neológica a través de continuas campañas de normalización

[5] El Dr. Ramón d'Andrés Díaz (Universidá d'Uviéu y ALLA), autor de varios artículos sobre la neología asturiana, tiene recopilada una rica bibliografía sobre la materia en los periodos anterior y posterior a la publicación del DALLA. Parte de ella está recogida en el apartado 6 («Bibliografía») del presente trabajo.

[6] La Dra. Ana Mª Cano González, presidenta de la ALLA y catedrática de Filología Románica en la Universidá d'Uviéu, ha descrito en diversos artículos los principios que guiaron la creación del DALLA. Destacan las referencias *Cano: 1993, 1995, 1996 y 2002*, recogidas en detalle en el apartado 6 («Bibliografía») del presente trabajo.

[7] *Surdimientu* (*surgimiento*) es el nombre con que se conoce al movimiento de dignificación de la lengua y la cultura asturianas iniciado en los últimos años de la dictadura franquista (1974) en torno a un grupo de universitarios y autores denominado Conceyu Bable, germen del núcleo fundador de la ALLA. Sus principios, sometidos a una primera redefinición a finales de los ochenta y actualmente en fase de revisión por una hornada de autores noveles, han sido el eje vertebrador de los treinta años de reivindicación asturianista en democracia y, claro está, de todo el desarrollo interno que ha experimentado la *llingua*.

[8] Para más información: http://www.politicallinguistica.org.

dirigidas a fomentar el uso del asturiano en áreas como la hostelería, el comercio o la educación infantil.

El ámbito académico es otra fuente importante de neología y de reflexión neológica. La iniciativa aquí partiría sobre todo de la Facultá de Filoloxía de la Universidá d'Uviéu, tanto desde su vinculación directa a la ALLA como por parte del trabajo independiente de grupos y profesores como el Seminariu de Filoloxía y el Dr. Ramón d'Andrés Díaz, autor de abundantes artículos y ponencias sobre la materia y de la colección *Cuestiones d'asturianu normativu*[10] (D'Andrés: 2001-2003).

Junto a la filología y la lingüística, donde se ha creado una rica neología en los últimos años, ha habido avances en otras áreas de las ciencias sociales y humanas (pedagogía, historia, antropología, sociología…) e incluso de las ciencias naturales y la técnica, donde destaca la obra colectiva *La ciencia, n'otres pallabres* (Viejo: 2008), coordinada por el Dr. Viejo y publicada precisamente para divulgar la neología científica en asturiano.

Las limitaciones que impone la administración autonómica, tanto dentro de su propio medio, la Radiotelevisión del Principado de Asturias (RTPA)[11], como fuera de él[12], y la presencia más que simbólica del asturiano en la prensa escrita han hecho de la prensa digital el principal medio de comunicación en lengua asturiana (con la notable excepción del semanario impreso *Les Noticies*, con doce años de existencia ininterrumpida).

La tarea neológica de la prensa digital asturiana es intensa, por más que haya un predominio palpable de contenidos locales y lo internacional —ámbito, por naturaleza, más dado a la creación neológica— tenga en ocasiones una presencia anecdótica. Esa capacidad de creación, unida a su carácter participativo y a su notable difusión, convierte a los medios digitales en la fuente más sólida para analizar la neología en *llingua asturiana*[13].

Para finalizar esta parte, cabe señalar la tarea neológica abordada por diversos colectivos ciudadanos, como el AXA (Aconceyamientu de Xuristes pol Asturianu)[14], en el ámbito jurídico, la iniciativa Softastur[15] de software libre, en

[9] Para más información sobre la oficina de Xixón: http://www.gijon.es/Contenido.aspx?id=18591&zona=2&leng=ast.

[10] Esta colección surgió precisamente de una iniciativa muy *2.0*: los comentarios con los que el Dr. D'Andrés iba respondiendo a las dudas que los usuarios le planteaban desde la lista de distribución Llingua-list (activa desde el año 2000), que él mismo moderaba.

[11] Pese a las trabas, los programas en lengua asturiana, relegados a franjas horarias tan poco atractivas como el mediodía del sábado, han sido líderes de audiencia en la RTPA desde que empezaron a emitirse en el año 2007, habiendo momentos incluso de liderazgo absoluto de audiencia frente a las grandes cadenas públicas y privadas de carácter estatal. Para más información sobre la programación en asturiano de la RTPA: http://www.estayadasturianu.es.

[12] Me refiero aquí a los procesos legales iniciados por el gobierno autonómico contra la emisora de radio en asturiano Radio Sele y la cadena de televisión TeleAsturias y que derivaron, en el caso de la emisora, en un cambio de frecuencia forzado (2006) y, en el de la cadena de televisión, en su cierre definitivo (2007).

[13] En el apartado 2.1 («Corpus de extracción: *Asturnews*») se explicarán con más detalle las características principales de la prensa digital asturiana.

[14] Para más información: http://www.xuristes.as.

[15] Para más información: http://wiki.altuxa.com/softastur/Portada.

informática, y la versión asturiana de la *Wikipedia* (*Uiquipedia*)[16], en cuanto a divulgación científica general. Existe incluso una lista de distribución específica sobre neología, *Léxicu internacional*, en la que participan algunos responsables de oficinas municipales de normalización. La progresiva generalización del concepto de Red social o Web 2.0[17] ayudará a que estas iniciativas se multipliquen y puedan ganar en repercusión.

Neología

El hecho de que la lengua catalana ocupe una posición de liderazgo en la investigación neológica dentro de la Romania —en el ámbito hispánico, incluso, con preeminencia sobre el castellano— responde, sin duda, a la experiencia, el compromiso, el rigor científico y el trabajo continuado de grupos como el OBNEO y el LATEL, ambos del IULA (Universitat Pompeu Fabra)[18], y de instituciones como el Institut d'Estudis Catalans y TermCat[19]. Tal es así que, en la investigación neológica actual, es prácticamente imposible obviar la experiencia catalana; algo que, en el caso de una lengua románica pequeña y minorizada como el asturiano cobra especial sentido.

Para el trabajo, en concreto, me he centrado en la labor metodológica realizada por el OBNEO, plasmada en la obra *Llengua catalana i neologia*, que tomo como referencia e inspiración prioritaria (Observatori de Neologia: 2004). Esto es cierto hasta tal punto que, incluso, podría decirse que mi meta última sería favorecer la creación en Asturies de un grupo semejante al OBNEO, con capacidad para ofrecer, salvando las distancias, unos resultados equiparables.

Del texto del OBNEO me interesan tanto la metodología que refleja como la toma de posición que representa. En cuanto al primer aspecto, la sistematicidad de los análisis recogidos y la coherencia con que se justifica la categorización propuesta han hecho de la clasificación de neologismos del OBNEO una suerte de *estándar* dentro del ámbito románico, válido para la realización de estudios comparados, interlingüísticos. Así, observatorios de peso, como el italiano, el portugués o el gallego, ya la han adaptado a su propia realidad. En mi caso, trataré de aplicarla con la máxima integridad posible a la neología asturiana.

El otro aspecto, de más calado, queda plasmado en la introducción que hace la Dra. Mª Teresa Cabré Castellví a la obra, con el elocuente título de «La importància de la neologia per al desenvolupament sostenible de la llengua catalana». En sus reflexiones habla del protagonismo de la neología a la hora de dotar a las lenguas minorizadas de los recursos de modernidad, flexibilidad e

[16] Según datos de la propia *Uiquipedia*, en agosto del 2007, la edición asturiana de la enciclopedia libre (activa desde el 2004) ocupaba el puesto número 77 en número de artículos (11.613), un lugar muy destacado teniendo en cuenta que el total de ediciones es de 262. La URL de la *Uiquipedia* es: http://ast.wikipedia.org.

[17] Para una definición *wiki* de este concepto: http://es.wikipedia.org/wiki/Web_2.0.

[18] Prueba de esta pujanza fue el éxito del Congrés Internacional de Neologia (CINEO 2008), celebrado en Barcelona el pasado mes de mayo y organizado por el OBNEO. Pensado como un congreso sobre neología románica, atrajo a especialistas de la práctica totalidad de la Romania —lenguas minoritarias incluidas— e incluso de otros dominios lingüísticos.
Para más información: http://www.iula.upf.edu (IULA), http://www.iula.upf.edu/obneo (OBNEO) y http://www.iula.upf.edu/latel (LATEL).

[19] Para más información: http://www.iec.cat y http://www.termcat.cat.

intercambio, necesarios para garantizar su supervivencia. Esta función planificadora de la neología, conceptualizada en la tradición *socioterminológica*[20] quebequesa y desarrollada hasta su máxima expresión en la experiencia catalana, sería, precisamente, la que buscaría desencadenar en la comunidad *asturfalante* con mis trabajos de sondeo y prospección.

Ingeniería lingüística

El aspecto técnico constituye el tercer punto de apoyo del sondeo, si bien es cierto que, de momento, su función es ante todo vehicular. El protagonismo se lo reservo para la fase de mi investigación de doctorado, la de la *prospección*.

La fuente de inspiración en cuanto a ingeniería lingüística está principalmente en las soluciones técnicas y estratégicas desarrolladas para el OBNEO por el LATEL y su investigador principal, el Dr. Lluís de Yzaguirre i Maura (Alonso: 2002 y De Yzaguirre: 2000a, 2000b, 2000c y 2001b), director de este trabajo. La intención final (fase de prospección) es aprovechar las estrategias de desambiguación, lematización y clasificación semiautomática de neologismos desarrolladas por el LATEL y adaptarlas a la realidad de la lengua asturiana, siempre desde una perspectiva panlatina[21].

En cuanto al objetivo inmediato de este sondeo, cuento con los consejos y la experiencia en programación en Perl de mi director, y, muy especialmente, con la idea de fondo de su concepto de *lingware neutral* (De Yzaguirre: 2001a), con el que ha querido señalar la necesidad de que los recursos de tecnología lingüística sean independientes de las lenguas de destino, como manera de evitar la exclusión que sufren las más pequeñas y minorizadas.

Asimismo, para procesos concretos del trabajo, como la extracción y explotación de los corpus o la obtención, búsqueda y detección de tipos generales de neologismos, he recurrido a herramientas públicas, disponibles en el *mercado* y de eficacia probada en proyectos semejantes: WGet[22], para la descarga automática de los corpus de extracción y de exclusión a partir de los sitios web de *Asturnews* y de la ALLA, respectivamente; Simple Concordance Program (SCP, de Alan Reed)[23], para la explotación de los corpus y la obtención de estadísticas y listados de léxico, y Search and Replace (de Funduc Software)[24], para las tareas masivas de búsqueda, corrección y eliminación de cadenas de texto, aspecto muy importante por los problemas de codificación detectados en las páginas web.

---

[20] En este punto abuso conscientemente del concepto de *socioterminología*, que utilizo de manera parcial y anacrónica para insistir en el lado más *socio* de mi trabajo. En realidad, el concepto no surgió como tal hasta los años ochenta, tras años de experiencia planificadora en Quebec (Gaudin: 2007).

[21] El objetivo será reciclar y adaptar estrategias desarrolladas para otras lenguas neolatinas, como el lematizador PALIC y el desambiguador AMBILIC (castellano y catalán), e integrar al asturiano en proyectos de alcance panrománico, como el de la *Taula panllatina de formants cultes*, elaborado por el IULA para Realiter (Red Panlatina de Terminología).
Para más información:
http://www.iula.upf.edu/recurs01es.htm y http://morgana.upf.es/cpt/formants/tf_i.htm.

[22] Versión empleada: WGet 1.11.4r1 (precompilada para Mac OSX).

[23] Versión empleada: SCP 4.0.9 (build 11).

[24] Versión empleada: Search and Replace 3.7 (para Windows).

Por otro lado, las tareas de programación, de las que se hablará con más detalle en el apartado 3 del trabajo («Metodología»), se han realizado íntegramente en Perl[25], tanto por creación directa como por adaptación de secuencias de código ya existentes. Perl es el lenguaje de programación favorito entre los especialistas en ingeniería lingüística por su flexibilidad y por la gran cantidad de recursos que hay disponibles a través de plataformas como CPAN[26].

Finalizado el sondeo, y como tarea previa que orientará la futura fase de prospección, tengo previsto explorar las posibilidades que ofrecen los interesantísimos trabajos de dos investigadores vinculados al IULA, Rogelio Nazar y el Dr. Maarten Janssen.

Nazar es el autor de la herramienta Jaguar[27], empleada para la explotación estadística de corpus por medio de un conjunto variado de estrategias (concordancias, recuentos de enegramas, extracción de colocaciones, obtención de medidas de asociación, distribución y similitud…). Me interesa particularmente su componente de representación gráfica de mapas conceptuales, que considero muy útil para la detección afinada de neologismos semánticos, aspecto especialmente complejo y difícil de tratar desde una perspectiva computacional (Nazar: 2007a, 2007b y 2008).

El Dr. Janssen, por su parte, es el autor de NeoTrack[28], una plataforma web para la detección semiautomática de neologismos en corpus electrónicos desarrollada para el Instituto de Linguística Teórica e Computacional (ILTEC) de Portugal (Janssen: 2005). Se aplica actualmente sobre una base de datos en lengua portuguesa, aunque con la intención de llevarla a otras lenguas cercanas. Su explotación requiere la lematización de los corpus de trabajo, lo que, en mi caso, podría retrasar una aplicación directa. Aun así, su carácter interactivo y las potentes herramientas que incorpora la convierten en una aplicación muy atractiva a efectos de planificación neológica, esencial para el caso asturiano[29].

Un último referente técnico lo constituiría BwanaNet[30], programa para la explotación del corpus técnico del IULA basado en el procesador de búsquedas en corpus (CQP) de la Universidad de Stuttgart. En este caso, se tendrá muy en cuenta su formato, considerado estándar en el sector, para la lematización y la explotación de los corpus generados durante este sondeo.

[25] Versión empleada: Perl v5.8.8 (para Mac OSX 10.5.4).
[26] Comprehensive Perl Archive Network: http://www.cpan.org.
[27] Para más información: http://www.iula.upf.edu/recurs01es.htm.
[28] Para más información: http://www.iltec.pt/pdf/wpapers/2005-mjanssen-oslin.pdf.
[29] Un valor añadido es el éxito que ha tenido esta plataforma en Portugal, cuya lengua y cultura ejercen una apreciable influencia sobre los círculos asturianistas.
[30] Para más información: http://bwananet.iula.upf.edu.

1.3 Marco teórico

El marco teórico de mi trabajo se define a partir de cuatro parámetros, dos metodológicos y dos relativos al objeto de estudio. El primero responde a la adopción de los principios y supuestos de la lingüística de corpus. El segundo se centra en la aplicación práctica al campo de la neología de procedimientos de trabajo y estrategias de ingeniería lingüística. Los otros dos, finalmente, surgen del doble objeto de estudio del sondeo, la neología como tal y su aplicación a lengua asturiana[31].

A continuación, iré desgranando las asunciones y las elecciones tomadas en cada uno de los parámetros con el objeto de perfilar el *envoltorio teórico* que rodeará a mi sondeo.

La opción, en primer lugar, de un corpus de uso real y actual como fuente para los datos del estudio implica automáticamente un posicionamiento en el eje empiricista, de la actuación (*performance*), opuesto —en principio— al eje racionalista, de la competencia (*competence*). Es la clásica disputa entre la lingüística de corpus y las visiones estructuralista y, sobre todo, chomskiana del sistema de la lengua, sobre la que ya se ha escrito abundantemente (McEnery y Wilson: 1996) y de la que no corresponde aquí hablar.

Sea como fuere, este posicionamiento no supone una toma de partida en la disputa sobre la esencia de la lengua: simplemente es una *conditio sine qua non* para un trabajo de este tipo. En ese sentido, el empleo de recursos de ingeniería lingüística también estaría teóricamente condicionado. No en vano, gran parte del éxito que han tenido las prácticas de ingeniería lingüística en las últimas décadas viene de su estrecha relación y complicidad con los logros de la lingüística de corpus.

En ambos casos, pues, asumo el papel de *investigador aséptico*, que, prescindiendo de la introspección, toma datos reales, los observa y los somete a análisis mediante un conjunto de estrategias ya validadas.

En lo que a la neología se refiere, adopto la perspectiva del *observador externo*, que la ve como un índice muy preciso para medir la vitalidad interna de una lengua, en este caso minorizada. El objeto aquí, como ya está dicho, sería elaborar, sobre la base metodológica del OBNEO, un producto que pueda servir como germen para posteriores actividades —institucionales o no— de planificación lingüística, a través de una fase intermedia de reflexión y de *veille néologique* desde un observatorio propiamente dicho. No tienen cabida en mi trabajo, por tanto, otras visiones de la neología, como la de su estudio interno, la lexicográfica o la centrada en el proceso de la innovación léxica.

El encuadre teórico que adopto respecto a la materia de mi estudio, la lengua asturiana, es doble: relacional en lo externo y funcional en lo interno. Relacional porque pretendo salir del discurso endogámico de la supervivencia y la

[31] En este último aspecto, desde un sesgo marcadamente sociolingüístico (planificación lingüística).

reivindicación para, mediante el uso de una metodología ya generalizada, permitir una comparación de los mecanismos neológicos del asturiano con los de otras lenguas hermanas (en lo románico y en lo minorizado). Habría en mi trabajo, pues, una perspectiva de fondo panlatina y *panminoritaria*.

El otro aspecto, prioritario, no hace sino seguir la línea argumental iniciada al comienzo de este apartado, cuyo objetivo es situar el sondeo dentro del paradigma funcional del lenguaje. Leyendo precisamente la siguiente descripción que de este paradigma hace uno de sus máximos exponentes, Simon C. Dik, quedan patentes muchos de los puntos de partida teóricos de mi trabajo:

> *In the functional paradigm a language is in the first place conceptualized as an instrument of social interaction among human beings, used with the intention of establishing communicative relationships. Within this paradigm one attempts to reveal the instrumentality of language with respect to what people do and achieve with it in social interaction. A natural language, in other words, is seen as an integrated part of the communicative competence of the natural language user.*
>
> DIK, S.C. (1989): *The Theory of Functional Grammar (Part I: The Structure of the Clause)*

En efecto, con mi estudio de la neología en asturiano, detectada por medio de herramientas y estrategias computacionales a partir de un corpus de uso real, busco medir su capacidad como lengua para servir de instrumento en las interacciones sociales de la sociedad moderna asturiana, es decir, su viabilidad como lengua de futuro.

1.4 Objetivos

El objetivo central de este proyecto es realizar una primera aproximación a la cuestión de la neología en lengua asturiana —*sondeo*, por tanto— a partir de un corpus de uso real (con textos periodísticos actuales), mediante el filtrado con otro corpus autorizado (de exclusión) y con el aprovechamiento y la adaptación de herramientas y estrategias tecnológicas que han demostrado su eficacia en otras lenguas del ámbito románico.

En paralelo al objetivo central, me planteo otros tres objetivos secundarios o indirectos:

1. Ahondar en la función sociolingüística de la neología, poniendo de manifiesto su valor y utilidad para lenguas como el asturiano, no sólo minoritarias sino también minorizadas y escasas de recursos.

2. Abrir nuevas vías en el debate sobre la supervivencia del asturiano, demasiado absorbido, quizá, por la cuestión reivindicativa, con la intención final de ayudar a acercarlo a las líneas de reflexión más pujantes actualmente en la Romania.

3. Realizar una pequeña aportación al imparable proceso colectivo por el que la comunidad asturianista está intentando dotar a la *llingua* de nuevos recursos y herramientas con los que impulsar su normalización.

Y, por supuesto, como ya está dicho, la meta última de mi trabajo sería sentar una base sobre la que construir un observatorio neológico mínimo, una atalaya de la llingua —virtual incluso— que sirviera no sólo para conseguir resultados y avances equiparables a los de los observatorios que ya trabajan en las lenguas románicas mayores, sino también para permitir al asturiano unirse a su rico debate actual y favorecer un diálogo práctico y constructivo con otras realidades.

2. Corpus de trabajo

Los corpus, de extracción (*Asturnews*) y de exclusión (*Diccionariu de la llingua asturiana*, DALLA), ocupan la parte central de mi trabajo. En ambos casos, proceden de fuentes digitales escritas exclusivamente en asturiano y suficientemente autorizadas: *Asturnews* como uno de los principales medios en lengua asturiana y el DALLA como texto clave del proceso de normativización de la *llingua* y obra cumbre de su máxima institución, la Academia de la Llingua Asturiana (ALLA).

La obtención de ambos recursos ha sido posible gracias a la confianza depositada en mí por el Dr. Próspero Morán López, periodista, académico correspondiente de la ALLA, profesor asociado de la Universidad de Valladolid y fundador y máximo responsable de *Asturnews*, y la Dra. Ana Mª Cano González, romanista, catedrática de Filoloxía Románica y ex decana de la Facultá de Filoloxía de la Universidá d'Uviéu y presidenta de la ALLA. Su disponibilidad y afán de colaboración con mi proyecto son una muestra más de su compromiso y entrega total con la *llingua asturiana*.

La composición de los corpus se ha realizado mediante descarga directa, con el programa WGet, de un conjunto fijo de documentos desde los sitios web de *Asturnews* y la ALLA, a saber:

- http://www.asturnews.com
- http://www.academiadelallingua.com

Para su tratamiento, una vez descargados los documentos, he utilizado un conjunto de aplicaciones Perl desarrolladas *ad hoc* y el programa Search and Replace. Para la explotación, he recurrido al programa especializado SCP.

Todos los procedimientos empleados quedan descritos con detalle en el siguiente capítulo («Metodología»).

2.1 Corpus de extracción: *Asturnews*

Activo en su formato actual desde enero del 2004[32], este *diariu cultural asturianu* es, junto con *Asturies.com*[33] y la edición digital de *Les Noticies*[34], uno de los grandes de la prensa digital en asturiano.

Lo de *grande*, por extraño que parezca, no es ni exagerado ni irónico: *Asturnews* recibe de media cerca de 300.000 visitas al mes[35], un volumen nada desdeñable teniendo en cuenta que —como ya comentaba en el «Prólogo»— su público potencial se situaría en torno a los 240.000 lectores, sobre una población *asturfalante* de entre 350.000 y 500.000 personas[36]. De acuerdo con su director, estas cifras le permiten al diario no vivir exclusivamente de las subvenciones públicas y recibir ingresos por publicidad.

Por generación de contenidos, *Asturnews* ocupa también un lugar privilegiado en la prensa digital en asturiano. Así, aparte de los artículos de opinión, el diario produce, desde la implantación de un nuevo diseño en el 2007, más de 400 noticias al mes, de ámbito sobre todo local (la escasez de contenidos de ámbito internacional es una constante en todos los medios asturianos) y temática dominantemente social y cultural. Basta con ver sus secciones: «Llingua y Sociedá», «Educación», «Música», «Lliteratura», «Cine», «Teatru», «Gastronomía», «Artes Plástiques», «Comunicación», «Opinión» y «Collaboraciones».

Por otro lado, el hecho de que el diario se edite con una licencia *copyleft*[37] ayuda a darle más difusión: muchas de sus noticias son reproducidas en blogs, sitios

[32] En este punto, utilizaré datos suministrados por el propio director de *Asturnews*. Proceden de su contribución para un monográfico sobre el periodismo en las lenguas minoritarias de España que será publicado en unos pocos meses por la Universidad del País Vasco.
En cuanto a su historia, cabe decir que, antes de su etapa actual, *Asturnews* vivió otra *embrionaria*, comprendida entre los años 1997 y 2000. Ya desde el principio, este diario nació con la vocación de «tapar el furacu que, nel ámbitu cultural, amosaba la rede Internet» y con el objetivo de «informar a los lectores asturfalantes de tol mundu, con rigor y en tiempu real, de l'actualidá cultural asturiana».

[33] *Asturies.com*, fundado en 1997 por la firma Araz y dirigido actualmente por Fernando de la Fuente Trabanco, es el diario decano de la prensa digital asturiana. Es, en realidad, parte de una plataforma mayor, *Asturia Activa*, a la que también pertenecen el buscador *Úlos*, la tienda electrónica *Astursho*p, la guía de ocio *Asturia Activa*, la web de turismo *Asturtravel* y el canal *Asturianía*, dedicado a la emigración asturiana. Asimismo, el diario tiene dos secciones multimedia, *Asturies.com TV* y *Asturies.com Radio*. Hasta el año pasado, además, contenía una sección muy popular de foros.
Para más información: http://www.asturies.com.

[34] Fundado por la asociación cultural Ámbitu (posteriormente, Publicaciones Ámbitu), el semanario *Les Noticies* es, por fortuna, la excepción de la prensa escrita asturiana. Su presentación, en 1996, se convirtió en su día en todo un acontecimiento, ya que vino a romper una tradición de 80 años sin prensa en asturiano (su antecesor fue el semanario regionalista *Ixuxú*, de principios del s. XX).
La edición digital arrancó en el 2007 y actualmente está a cargo de Xuan Bello, el escritor en asturiano más reconocido dentro y fuera de Asturies (ya ha sido traducido, incluso, a castellano y catalán).
Para más información: http://www.lesnoticies.com.

[35] Según cifras del propio diario obtenidas mediante la herramienta Google Analytics a finales del 2007: http://www.asturnews.com/index2008.php?idn=5472&hemeroteca=1.

[36] Las cifras de hablantes oscilan entre el cálculo conservador de 350.000 del Dr. D'Andrés y los 500.000, equivalentes al 49 % de asturianos que afirma entender y hablar el asturiano de acuerdo con el *II Estudio sociolingüístico de Asturias (2002)*, de Francisco J. Llera Ramo y la ALLA.
La población total de Asturies es de 1.076.635 habitantes, según el último censo disponible, del 2005 (fuente: INE).

[37] El texto completo de la licencia se encuentra en: http://www.asturnews.com/popup_condiciones.php.

web e incluso en otros medios digitales. De esta forma, acaba funcionando como una especie agencia de noticias en asturiano y ve multiplicado su alcance.

En cuanto al equipo redactor, está compuesto en la actualidad por tres personas, incluido el director, con el apoyo de otras dos más para cuestiones de maquetación y programación multimedia. En secciones concretas, como las de opinión y música, el diario también cuenta con colaboradores externos, aunque de manera ocasional. Por último estaría Alberto Vázquez, autor de la popular viñeta semanal protagonizada por la *paisana* Felecidá Comerón. El diario cuenta con una pequeña guía de estilo redactada por su director.

Como se ve, tanto por difusión como por volumen, variedad y renovación de contenidos, estaría suficientemente justificada la elección de *Asturnews* como fuente para la detección de neología frente a otros medios digitales escritos en asturiano (por ejemplo, *Asturies.com*, *Les Noticies*, el sitio *Estaya d'asturianu* de la RTPA —Radiotelevisión del Principáu d'Asturies— o el canal de lengua asturiana de Europa Press…)[38]. La licencia *copyleft* con que funciona este medio, que autoriza la difusión gratuita de contenidos siempre que se cite debidamente su procedencia, el prestigio personal y académico de su director y, sobre todo, su disponibilidad y entusiasmo a la hora de involucrarse en el proyecto de sondeo neológico son otros de los motivos que han sustentado mi elección.

El material del corpus de extracción está compuesto por todas las noticias aparecidas en *Asturnews* en los primeros cuatro años y medio de su actual etapa, en concreto, entre el 24 de enero del 2004 y el 30 de junio del 2008. Son, en total, más de 6.800 noticias, las comprendidas entre las siguientes URL:

a) Para la temporada 2004-2007:

http://www.asturnews.com/index2008.php?idn=**0001**&hemeroteca=1

…

http://www.asturnews.com/index2008.php?idn=**5664**&hemeroteca=1

b) Para la temporada 2008:

http://www.asturnews.com/index2008.php?idn=**5665**

…

http://www.asturnews.com/index2008.php?idn=**6824**

[38] Las direcciones de los medios aún no citados son:
http://www.estayadasturianu.es y http://www.europapress.es/asturies/asturianu.

Asimismo, se han incluido unos 100 *posts* de opinión, es decir, todos los publicados desde que se inició la sección (24 de enero del 2006) hasta la primera mitad del 2008 (29 de junio). Las URL en cuestión serían las comprendidas entre:

http://opinion.asturnews.com/?p=**9**

…

http://opinion.asturnews.com/?p=**121**

Una vez descargadas las URL, tuvieron que pasar por un proceso de *limpieza* para permitir su explotación como corpus. Este proceso se realizó en varias etapas: a) conversión de los archivos HTML a formato .txt; b) eliminación de los códigos HTML y los textos irrelevantes (enlaces a otras noticias y secciones, textos fijos…); c) conversión de los errores de codificación detectados[39]; d) separación gráfica de las formas apostrofadas[40] y los clíticos de dativo (*-y*, *-ys*, *-yos*, *-ylu*, *-yla*, *-ylo*, *-yoslo*…); e) uniformización tipográfica de las consonantes dialectales (*hache aspirada* y *che vaqueira*)[41] y f) integración de todos los archivos en uno solo.

Para estos pasos intermedios, que se describen con más detalle en el apartado 3 («Metodología»), empleé varias secuencias escritas en Perl y el programa de búsqueda y sustitución en bloque de textos Search and Replace.

Con los casi 7.000 archivos procesados en el corpus, obtuve un volumen de texto superior a 1.500.000 palabras[42], suficiente para garantizar la representatividad de la muestra (es probablemente, uno de los corpus más grandes recopilados hasta la fecha en asturiano)…, y excesivo para su manipulación[43].

[39] Estos errores procedían sobre todo de la sección de opinión, cuyas páginas presentan una estructura de código fuente distinta a la del resto de secciones y reciben más comentarios libres por parte de los lectores. Estos comentarios no siempre son en asturiano: se han detectado también en castellano, catalán, gallego, gallegoasturiano y portugués (todos ellos se han conservado íntegros en el corpus).

[40] Las palabras que apostrofan en asturiano son: *en*, *la*, *de*, *el*, *pa*, *que*, *me*, *te*, *se* (todas hacia delante, excepto *el*, que puede apostrofar hacia delante o hacia atrás: *L'aire'l monte ye perbono*).

[41] Son las llamadas *consonantes sopuntiaes* (*h* y *ll* con un punto debajo de cada letra), auténtico quebradero de cabeza tipográfico del asturiano moderno, ya que faltan en la mayoría de fuentes de los procesadores de texto más habituales.
Están pensadas para representar dos grupos de sonidos, en principio, de carácter dialectal: las aspiradas y velares del asturiano oriental, en el caso de la *h*, y los sonidos entre retroflexos y palatales con que, en zonas del asturiano central y, especialmente, en el asturiano occidental, se pronuncian *grosso modo* las *ll* del resto del dominio (idénticas éstas a las laterales palatales catalanas, portuguesas, italianas…).
En cuanto a las aspiradas, los usuarios suelen sustituirlas por *h.* o por *j*, mientras que las ches vaqueiras se convierten en *l.l* (el punto debajo y sólo en el medio) o *!!* (dos signos exclamativos de cierre).

[42] Exactamente, 1.534.518 palabras, según el sistema de recuento automático de Microsoft Word (uno de los más comúnmente empleados en la práctica de profesiones que *viven de sus palabras*, como los traductores o los periodistas).

[43] La manipulación tanto de los archivos como de las palabras de ambos corpus ha sido la causa de los principales problemas técnicos que ha conllevado realizar este trabajo. Así, por ejemplo, el procesamiento de los archivos ha causado problemas inesperados con los códigos de Perl, en su ejecución tanto sobre Unix como sobre DOS (el trabajo lo he realizado sobre dos plataformas, Mac y Windows), mientras que el volumen de palabras ha limitado el funcionamiento tanto de la herramienta principal para la explotación de los corpus, SCP, como de otras empleadas en tareas accesorias (Microsoft Word y Microsoft Excel).

Con posterioridad al sondeo neológico, mi intención es que el corpus de extracción pase a estar a la disposición de otros investigadores y del público en general —siempre con la autorización de *Asturnews*—. Para ello, pretendo integrarlo en el proyecto de corpus electrónico para el asturiano de la Universidá d'Uviéu (Eslema) y, en paralelo, colocarlo en un sitio web que permita una explotación basada en el procesador CQP e inspirada en la que realiza el programa BwanaNet (IULA), para lo que será necesaria su lematización.

De momento, he incluido una copia digital de ambos corpus (extracción y exclusión) como anexo a este sondeo. En el apartado «Anexos» se indican los nombres de archivo exactos, junto con una ruta web para la descarga gratuita del programa SCP.

2.2 Corpus lexicográfico de exclusión: DALLA

El pasado mes de marzo, la ALLA presentaba *en sociedad* la versión digital de su máxima obra, el *Diccionariu de la llingua asturiana*, publicado originalmente en el año 2000. La plataforma elegida fue su propio sitio web y el sistema de consulta, el más sencillo: consulta directa y libre, sin necesidad de registro o de identificación de usuarios. En paralelo, la Academia decidió colgar para cada consulta dos enlaces directos a sus otras grandes obras normativizadoras, la *Gramática* y la *Ortografía* oficiales. Con ello, vino a satisfacer el deseo de muchos profesionales y estudiosos de la *llingua* y supo dar un paso irrevocable hacia la difusión y generación de recursos técnicos en asturiano.

En realidad, la publicación *online* del DALLA supuso la culminación de un proceso de apertura que se inició en el 2004, con la salida al mercado, primero mediante pago y luego de forma gratuita, por descarga, del Iguador, el primer corrector ortográfico para asturiano.

El Iguador es una sencilla pero potente aplicación que emplea una base de datos y un conjunto de reglas para generar, en principio, todas las formas posibles de las más de 51.000 entradas del DALLA. Concebido para llegar al máximo de usuarios, es decir, para funcionar sobre plataformas Windows, su diseño responde a los requisitos del *Catálogo de soluciones de Office*. Para ello fue necesario un acuerdo a tres bandas entre la ALLA, Microsoft y Araz, empresa responsable de su ejecución (y, por cierto, de la edición digital del DALLA), lo que, a la postre, invistió de prestigio al producto: algo nada baladí para una lengua que se mueve habitualmente en el desprestigio.

Tanto el DALLA como el Iguador suponen dos obras inmensas, habida cuenta de la precariedad que rodea al asturiano, excluido de canales adecuados de financiación por su inexplicable condición legal de lengua *no cooficial*. Son fruto del compromiso y la buena sintonía de la ALLA y la firma Araz. En este sentido, una de las metas indirectas del sondeo es generar materiales que puedan servir tanto a la ALLA como a Araz para el afinamiento de sus recursos actuales (por ejemplo, mediante la inclusión de neologismos en el Iguador) y la creación de otros nuevos.

Volviendo al corpus lexicográfico de exclusión, éste se basa en el texto completo de las **51.648 entradas** que componen el DALLA, obtenido, al igual que hice con el corpus de extracción, mediante descarga directa desde Internet con el programa WGet.

El intervalo de URL elegido para la descarga fue el siguiente:

http://www.academiadelallingua.com/diccionariu/index.php?cod=**00001**

…

http://www.academiadelallingua.com/diccionariu/index.php?cod=**51648**

O, lo que es lo mismo, todas las entradas del DALLA, desde a hasta zutrón.

Una vez hecha la descarga, tuve que someter los archivos a un procesamiento semejante al aplicado a los de *Asturnews* (conversión a texto, limpieza de código, unificación en un solo archivo, uniformización tipográfica…).

A diferencia del corpus de exclusión, en el corpus del DALLA no tuve problemas de codificación —ya que procedían de una misma fuente—, aunque sí tuve que someter el texto completo a otros tratamientos, como la eliminación de las abreviaturas lexicográficas y gramaticales[44]. Para ello, utilicé, una vez más, el programa Search and Replace.

En cuanto a volumen, el DALLA completo implicó la manipulación de casi 52.000 archivos, lo que causó no pocos problemas de procesamiento. En conjunto, resultó un cuerpo de texto de más de 900.000 palabras[45].

Aunque no sea exhaustivo, este conjunto se considera aceptablemente representativo, ya que el DALLA incluye ejemplos y fraseología, aparte de las propias entradas y definiciones. Una vez finalizado el sondeo, el catálogo de formas del DALLA será puesto a la disposición de la Academia para, entre otras finalidades, permitir la detección de errores y realizar estudios comparados a partir de su contraste con las formas posibles que genera el Iguador.

[44] He conservado, no obstante, una copia *en sucio* (con códigos HTML y etiquetas lexicográficas) de la edición digital del DALLA, ya que, aunque no sirva directamente para este sondeo, toda esta información estructurada puede resultar muy valiosa tanto para fases posteriores de mi trabajo como para otras investigaciones indirectas.

[45] La cifra exacta de palabras aquí fue de 907.448, también según el contador automático de Microsoft Word. En conjunto, ambos corpus suman 2.441.966 palabras.

3. Metodología

La palabra *sondeo* es toda una declaración de intenciones. En efecto, mi trabajo pretende tan sólo ser una primera toma de contacto con la neología en asturiano que permita diseñar estrategias y *vías de ataque* para la posterior *prospección*, para la que será necesaria una fase previa de lematización y de diseño y adaptación de herramientas específicas de detección y clasificación neológica.

En cuanto al sondeo en sí, ha implicado tres fases diferentes, cada una de ellas con su propio enfoque metodológico y subfases:

1) Preparación del corpus de extracción

    1. Descarga de archivos
    2. Limpieza y procesado de los textos
    3. Confección del corpus en SCP

2) Confección del corpus lexicográfica de exclusión

    1. Descarga del diccionario
    2. Limpieza y procesado de los textos
    3. Confección de un corpus en SCP y extracción de unidades léxicas
    4. Refinamiento de la lista lexicográfica de exclusión

3) Extracción y clasificación superficial de neologismos

    1. Aplicación de la lista de exclusión sobre el corpus de extracción
    2. Criba parcial de erratas y falsos neologismos
    3. Clasificación manual de los neologismos detectados

3.1 Preparación del corpus de extracción

El primer paso para la preparación del corpus de extracción, con todas las noticias de *Asturnews* de enero del 2004 a junio del 2008, consistió en la descarga de las casi 7.000 URL del sitio web del diario.

La descarga se efectuó mediante un comando de WGet. El programa, de distribución gratuita mediante una licencia de GNU, es muy popular para este tipo de tareas por su solidez y flexibilidad. Ahora bien, como trabaja desde la línea de comandos (sin interfaz gráfica), requirió de mi parte la adquisición de ciertos conocimientos previos para su correcta configuración[46].

Entre otras cuestiones, tuve que generar un archivo de texto con todas las URL de la descarga que, después, pasé a WGet. Para ello, preparé un sencillo código en Perl[47] en el que, a partir de las URL de origen deseadas (de la sección principal y de opinión de *Asturnews*[48]), generé el resto mediante un procedimiento de bucle, aprovechando que obedecían a sendas sucesiones (del 0001 al 6824 en el caso de la sección principal y del 9 al 121 en la de opinión[49]).

El comando que envié a WGet fue exactamente:

```
wget -x -i /Users/afcernuda/URL_asturnews.txt -E
```

En él, especificaba la descarga en un mismo directorio (`-x`) de todas las URL contenidas en un archivo de entrada (`-i`), generado antes mediante el código el Perl indicado y para el que tuve que señalar una ruta local (`/Users/…/URL_asturnews.txt`). Por último, tuve que especificar que el programa recompusiera los archivos descargados en su formato original (HTML), opción predeterminada en el comando `-E`.

La eliminación de los códigos HTML[50] y la extracción de la información válida para el análisis neológico (es decir, de los textos de los titulares, los cuerpos de las noticias, los pies de fotos e imágenes y los comentarios de los lectores[51]), absorbió una importante cantidad de tiempo. Podría decirse, de hecho, que planteaba un problema por encima de mis posibilidades.

---

[46] Utilicé como obra de referencia el manual en línea *GNU WGet 1.11.4*, disponible en: http://www.gnu.org/software/wget/manual/wget.html.

[47] Todos los programas preparados para el trabajo se adjuntarán de forma digital en los anexos del sondeo. La relación completa de archivos se recoge en el apartado «Anexos».
Como referencia para la programación en Perl utilicé los siguientes manuales, todos ellos de nivel básico:
- *Perl 2007* (de Erik Tjong Kim Sang), disponible en http://ifarm.nl/erikt/perl2007.
- *Perl pour les linguistes* (de Ludovic Tanguy y Nabil Hathout), disponible en http://perl.linguistes.free.fr.
- *Perl Manual* (de Larry Wall), disponible en http://www.math.utah.edu/docs/info/perl_toc.html.

[48] Para más detalles sobre las URL concretas, vid. apartado 2.1 («Corpus de extracción: Asturnews»).

[49] Tras la descarga, pude constatar que, en la sucesión, había unas pocas URL de opinión sin contenido.

[50] Los archivos descargados en HTML se conservaron aparte, por si podían ser necesarios más adelante.

[51] Para el corpus se descartaron tanto los contenidos fijos de las URL (secciones, textos legales…) como otros textos variables (enlaces a otras noticias).

Y es que ni WGet ni ninguna aplicación comercial accesible para un usuario medio como yo[52] permitían una limpieza sencilla de los códigos. Para solucionar mi problema, recurrí a la ayuda de un compañero informático, buen conocedor de Perl, Rubén Abad.

Tras estudiar a fondo la estructura de los códigos fuente de las URL de *Asturnews* (en sus dos formatos: secciones comunes y sección de opinión), determinamos que la mejor opción era adaptar un módulo de marcaje en Perl que, invocando a su vez a otro módulo sacado de la plataforma CPAN, permitía definir qué campos concretos de cada etiqueta HTML eran de interés. Los campos que no quedaran marcados como válidos serían eliminados y los que sí, se descargarían como texto puro en un conjunto de archivos .txt.

Tras varios ajustes en el módulo[53], finalmente se consiguió un conjunto de textos *limpios* de etiquetas de HTML. Todos ellos fueron posteriormente integrados en un solo archivo de texto a través de un sencillo comando `copy` en DOS.

En un primer análisis superficial del archivo generado, se detectaron abundantes problemas de codificación que, por fortuna, respondían a dos grandes tipos. Por un lado estaban los restos de códigos HTML textuales y numéricos empleados para representar caracteres especiales (por ejemplo, `&aacute;` o `&224;` en lugar de `á`). Para su eliminación, recurrí a la lista oficial de códigos del consorcio W3C para las lenguas de Europa occidental (bloque *Latín 1*)[54], la generé en un *script* de sustitución de Search and Replace, donde introduje los correspondientes caracteres, y apliqué una sustitución masiva (siguiendo el ejemplo, hice sustituir por `á` cada aparición de `&aacute;` o `&224;`).

Tras ello, pasé a los errores causados por conversiones entre sistemas de codificación (por ejemplo, de UTF o Unicode a ANSI: cadenas `Â'` en lugar del carácter `Ñ`, entre otros muchos). Aquí la conversión fue más complicada, ya que, prácticamente, tuve que ir deduciendo los caracteres uno por uno. En cuanto saqué una lista completa, apliqué una sustitución masiva en Search and Replace.

Por último, realicé varias sustituciones masivas destinadas a liberar de peso el archivo (sustitución de dobles espacios y de caracteres innecesarios) y a uniformizar las grafías del texto conforme a cuatro sencillas reglas, destinadas a facilitar la posterior lematización del corpus.

[52] Todas las tareas asociadas a la preparación del sondeo se han realizado con limitación de medios humanos (una persona con conocimientos elementales de programación, ayudada puntualmente por expertos) y técnicos (dos ordenadores personales trabajando por separado, en distintas etapas, y software o bien libre o bien fácilmente accesible). Aparte de los condicionantes de un trabajo de máster, se ha querido mostrar todo lo que se puede conseguir con un esfuerzo razonable. La idea, una vez más, es impulsar la acción coordinada de los *pequeños planificadores*.

[53] La primera limpieza dejó un corpus aproximado de 900.000 palabras. De su revisión manual, mediante muestras aleatorias de archivos, se elaboró una lista de errores que, posteriormente, fueron corregidos en el módulo de marcaje. Tras el ajuste y la comprobación mediante la misma muestra de archivos, se llegó al corpus final de 1.500.000, considerado como válido.

[54] Disponible en http://www.w3.org/MarkUp/html3/latin1.html.

Estas reglas fueron:

1) separar mediante espacios las formas apostrofadas:

   p. ej.: convertir `l'aire'l monte` en `l' aire 'l monte`

2) separar mediante espacios las formas clíticas de dativo:

   p. ej.: `dába-yoslo` *(se lo daba a ellos)* por `dába -yoslo`

3) unificar los casos de hache aspirada en un dígrafo común que, para evitar confusiones, no utilizara el carácter punto (`.`):

   p. ej.: `guah.e` *(chico)* por `guah·e`

4) unificar los casos de che vaqueira en otro dígrafo sin punto (`.`) ni punto en voladilla (`·`):

   p. ej.: `L.lena` (concejo en el sur de Asturies) por `L_lena`

Una vez editado gráficamente el archivo, procedí a crear un proyecto de corpus en SCP. Para ello, en primer lugar, tuve que definir un alfabeto propio para el asturiano, de modo que las palabras quedaran ordenadas adecuadamente. Siguiendo las opciones de SCP y las convenciones tipográficas establecidas, configuré el sistema para pasar por alto guiones y apóstrofos y para considerar *h·* y *l_l* como variantes de *h* y *ll*, respectivamente.

Tras estos ajustes, el corpus se generó sin mayor problema que una limitación propia del programa SCP: a la hora de extraer la lista de palabras del corpus, los resultados se truncaron en 32.000. Esto complicaba enormemente la tarea posterior de contrastado con el corpus de exclusión.

3.2 Confección del corpus lexicográfico de exclusión

La primera parte de esta fase, la descarga de las más de 50.000 URL del DALLA, se completó mediante un proceso idéntico al del punto anterior: creación de todas las URL en un archivo de texto mediante un pequeño programa generador en Perl (por fortuna, también seguían una sucesión, de la 00001 a la 51648) y envío de un comando de descarga parametrizado a través de WGet:

```
wget -r -x -i /Users/afcernuda/URL_DALLA.txt -R "*.pdf" -E
```

En este caso, apliqué por error un comando de recursividad (`-r`), que provocó la descarga, por seguimiento de enlaces, de unos 150 ficheros aislados[55]. Asimismo, esta vez intencionalmente, apliqué la opción (`-R "*.pdf"`), con la que evité que el programa se descargara los enlaces a los pdf con la Gramática y la Ortografía oficiales que incluye cada una de las entradas del diccionario.

Una vez descargadas todas las entradas, decidí conservar una copia con el código fuente original (especialmente sencillo y dedicado sobre todo a aplicar marcas tipográficas). Y es que, si observamos una entrada cualquiera, tal cual aparece en el diccionario:

> **llingua, la:** *sust.* Muérganu [allargáu y musculosu de la boca, que val pa saborgar los alimentos, p'ayudar a articular soníos]. **2** Sistema [humanu de comunicación, formáu por unidaes articulaes provistes de significáu]. **3** Sistema de comunicación [oral propiu d'un pueblu, d'una comunidá]. **4** Manera o estilu d'espresase de pallabra. *Al falar en públicu hai que cuidar la llingua*. **5** Vezu de falar más de la cuenta. *Tien muncha llingua. La llingua ye lo que lu pierde*. || **A media llingua**, pronunciando mal dellos soníos [falar, dicir daqué]. || **Comer la llingua'l gatu a daquién**, *fam.* *espresión que se diz cuando daquién calla, cuando nun fala. || **Dar (a) la llingua**, *fam.* falar [muncho y de siguío]. || **Echar la llingua a pacer**, falar de daqué que nun se debe. **2** Falar mal de [daquién]. 3 Falar muncho, charrar de contino. || **La llingua [te] ampolle**, *espresión que s'usa p'amosar cansanciu por aguantar a ún que fala muncho o más de la cuenta. || **Llingua afilada**, enclín a criticar a los demás. || **Llingua de fueu**, vezu de falar mal, de dicir cagamentos. || **Llingua de güe**, planta [de fueya ancho que da flores grandes y blanques, y un frutu asemeyáu a una panoya]. **2** *Phyllitis scolopendrium*, [tipu de] felechu [de fueya allargao y ovalao ensin divisiones]. […] || **Llingua d'oc**, occitanu, llingua d'Occitania. […] || **Soltar la llingua**, facer que [daquién] fale por demás. *El ron soltó-y la llingua*. […]

Puede apreciarse que contiene mucha información que, en fases posteriores del trabajo (prospección), podría resultar muy valiosa para extraer, por ejemplo: unidades léxicas (en negrita: *llingua*, *a media llingua*, *comer la llingua'l gatu*…), ejemplos de uso (en cursiva: *Al falar en público hai que cuidar la llingua*, *Tien muncha llingua*…), sentidos estructurados y, por tanto, clasificables en forma de base de datos (definiciones completas), sentidos amplios para la creación de marcos léxicos y la búsqueda de sinónimos (definiciones sin corchetes)…

[55] La eliminación de estos archivos fue sencilla: como se grabaron con un nombre distinto del resto (fueron generados por recursividad, no porque figuraran en el archivo de URL válidas), durante el proceso de *limpieza* de HTML simplemente no fueron pasados por el módulo de marcaje, restringido a las URL válidas.

De todas maneras, como ya he comentado, para el primer sondeo bastaba con extraer el texto puro de las entradas, por lo que se sometió al lote de archivos descargados del DALLA a un proceso semejante al descrito antes para los de *Asturnews*: aplicación del modulo de marcaje en una versión adaptada a los códigos de las entradas del DALLA, extracción de los textos válidos seleccionados y volcado en archivos .txt nuevos, eliminación de errores de conversión (mínimos en este caso) y de cadenas generadoras de *ruido* (dobles espacios, por ejemplo) y aplicación de la convención tipográfica definida.

Como aspecto exclusivo de este caso estuvo la eliminación, por medio del programa Search and Replace, de las cadenas empleadas para las marcas lexicográficas, lo que, indirectamente, permitió sacar algunas estadísticas interesantes con respecto al contenido del diccionario: así, por ejemplo, pude observar que más de la mitad de las entradas son sustantivos (cadena `sust.` eliminada en 27.619 entradas de un total de 51.648), que hay una cantidad notable de verbos y adjetivos (cadenas `ax.` y `v.`, eliminadas en 11.289 y 9.142 entradas, respectivamente), que el dialecto occidental está mucho más representado que el oriental (117 cadenas `occ.` frente a 9 `or.`) o que la marca del artículo (`art.`) es del todo innecesaria, ya que sólo aparece en una entrada (*el, la, lo*) y lo hace una sola vez[56]…

También fue necesario eliminar las numerosas cadenas con marcas de flexión que existían en las entradas (por ejemplo, `asturianu, -a, -o`) para evitar que afectaran a la detección de neologismos[57].

Tras la eliminación de estas cadenas, se generó el corpus en SCP, con los mismos criterios aplicados al corpus de *Asturnews*. Una vez más, se desbordó la capacidad de procesamiento del programa, que fue incapaz de generar una lista de palabras más allá de las 32.000.

[56] Además, la definición que se da es, literalmente: «Artículu».
[57] Estas cadenas serán útiles en la fase de prospección, ya que facilitarán la modelización de lemas.

3.3 Extracción y clasificación superficial de neologismos

Una vez refinados los corpus de extracción de neologismos (*Asturnews*) y de exclusión de formas ya existentes (DALLA), podía pasarse a la siguiente fase, la obtención de las unidades léxicas de cada conjunto para, mediante la siguiente operación de resta, obtener una lista de candidatos a neologismos:

extracción (*Asturnews*) – exclusión (DALLA) = candidatos

Lo que equivalía a decir que aquellas unidades del corpus de *Asturnews* que no figuraran en la lista del DALLA se considerarían candidatas a neologismos.

Como ya adelantaba en los puntos anteriores, el problema estaba en que con SCP sólo había conseguido sacar una lista truncada de 32.000 unidades léxicas por cada corpus.

El nuevo bloqueo se resolvió con ayuda del Dr. De Yzaguirre, que decidió reaprovechar un código Perl que había desarrollado para una aplicación anterior. Reajustamos el programa para que sacara un listado con todas las formas de cada corpus (distinguiendo entre mayúsculas y minúsculas) y las enumerara en sucesión. Para cada forma se añadió un recuento compuesto por cuatro cifras (de derecha a izquierda en el ejemplo):

1) apariciones exclusivamente en mayúsculas en el corpus de *Asturnews*

2) apariciones totales en el corpus de *Asturnews*

3) apariciones exclusivamente en mayúsculas en el corpus del DALLA

4) apariciones totales en el corpus del DALLA

Por ejemplo:

```
0    4717 0    514  llingua

1391 0    160  0    Llengua
```

Quiere decir que la forma *llingua* figura 514 veces en el DALLA (160 de las cuales exclusivamente en mayúsculas) y 4.717 en el corpus de *Asturnews* (1.391 de ellas, en mayúsculas)[58].

[58] El archivo de texto con todos estos resultados también se ha incluido en los anexos al sondeo.

Por tanto, para la detección superficial de neologismos[59] bastaría con buscar unidades que tengan el valor `0` en los dos últimos campos:

```
0     1     0     0     asturtzale

81    0     0     0     Asturianía
```

En este caso tenemos dos neologismos *de manual*: *asturtzale* (radical asturianista que recuerda a los *abertzales* vascos) y *Asturianía* (colectividad formada por todos los asturianos, con especial atención a los emigrados; distinto de *asturianía* con minúsculas, cualidad de asturiano, que no es neologismo).

Para sacar un listado final de neologismos superficiales, sería suficiente, por tanto, con eliminar todo el ruido de la lista de candidatos, es decir:

1) las unidades que, independientemente de que figuren en *Asturnews*, aparezcan al menos una vez en el DALLA

2) las unidades exclusivas de *Asturnews* que sean formas válidas flexionadas de otras presentes en el DALLA:

   `0     2     0     0     comunico` (del verbo *comunicar*)

3) las erratas, faltas de ortografía y caracteres sueltos

4) las formas gráficas que reflejan usos dialectales o coloquiales:

   `0     1     0     0     al_leranos` (en lugar de *ayeranos*, de *Ayer*)

   `0     23    0     0     ná` (en lugar de la correcta *nada*)

5) los nombres propios[60] (abundantísimos en el corpus de *Asturnews* al tratarse de un diario cultural)

6) las palabras procedentes de textos escritos en otros idiomas

7) restos de código HTML o de contenidos fijos de las URL

   …

[59] Me refiero a los neologismos detectables por su forma y, por tanto, procesables de forma automática. Quedan fuera de esta clase las categorías que no se pueden detectar por contraste de listados de unidades: la neología semántica (asignación de nuevos significados a unidades ya existentes, como *roblar*, que ha incorporado el sentido *firmar* al original de *cerrar un trato mediante un apretón de manos*) y la sintagmática (formación de nuevos significados por combinación de dos o más unidades, como *páxina Web*, con un significado totalmente diferente del de *páxina* y *Web* por separado).

[60] Los topónimos, pese a ser nombres propios, quedarían excluidos de esta lista de elementos de ruido, ya que ofrecen un índice muy interesante de la capacidad de adaptación e internacionalización de la *llingua*: *La Rioxa* (frente al oficial *La Rioja*), *Estaos Xuníos* (frente al oficial *Estaos Uníos*), *Bruseles*, *Abhasia*…

La operación, simplificada, sería como sigue:

candidatos – ruido = neologismos superficiales

Después, en teoría, para completar nuestra lista *sólo bastaría* añadir los neologismos no superficiales (semánticos y sintagmáticos). La operación, sin embargo, resulta especialmente compleja, tanto, que, de hecho, es uno de los temas centrales de los estudios neológicos actuales. Técnicamente, requiere estrategias mucho más avanzadas que la detección superficial, como, por ejemplo, la lematización, categorización y, por ende, desambiguación de unidades léxicas; la realización de análisis morfosintácticos o de roles semánticos en las estructuras sintagmáticas; la elaboración de cálculos estadísticos, en ocasiones, con representación gráfica[61], mediante complejos algoritmos relacionales…).

En suma, los neologismos no superficiales exigirían soluciones más allá de las previstas para el sondeo, de ahí que deba prescindir de ellos en esta fase, a excepción de unos pocos detectados manualmente y que considero de interés.

Para procesar el listado, por tanto, la técnica prevista era la de extracción del ruido. Una vez más, el volumen de palabras provocó serios problemas: así, aunque el código del Dr. De Yzaguirre logró generar correctamente un archivo con todas las unidades diferentes de ambos corpus (extracción + exclusión: en total, más de 2,4 millones de palabras), el volumen final fue de tal magnitud (153.924 unidades[62], en un archivo de texto de 2,81 MB) que desbordó la capacidad de tratamiento no sólo de SCP, sino también de programas comerciales y muy potentes como Microsoft Word y Excel.

La única forma viable que vi de manejar el listado de candidatos fue dividirlo a mano en archivos ordenados alfabéticamente (candidatos empezados por A, B, C…). Tras horas de trabajo, pude observar patrones de ruido comunes (nombres propios, caracteres extraños, palabras en castellano…), por lo que, combinando varias estrategias (sustituciones en bloque en Search and Replace, ordenación ascendente o descendente por número de apariciones con Word y Excel…), aún pude eliminar unos cuantos miles de formas.

Aun así, no conseguí procesar por completo los cientos de miles de formas generadas. En el proceso de prueba, no obstante, logré detectar varios neologismos interesantes y establecer dos pequeños listados parciales, uno cualitativo (por categorías de neologismos) y otro cuantitativo (por frecuencia de aparición), que se recogen en el siguiente capítulo («Resultados»).

[61] Como los realizados por la herramienta Jaguar de Rogelio Nazar (vid. apartado 1.2, «Antecedentes de estudio»).

[62] Distribuidas en 92.237 unidades dentro del corpus de *Asturnews* (23.719 mayúsculas y 68.518 minúsculas) y 82.907 en el corpus del DALLA (14.193 mayúsculas y 68.714 minúsculas). La diferencia entre el total de formas (153.924) y la suma del total de unidades de cada corpus (175.144) nos da un total de 21.220, que representa las unidades coincidentes en ambos. El volumen de formas por revisar se obtendrá, por tanto, de la diferencia entre el total de formas diferentes (~154.000) y el conjunto de unidades comunes (~21.000), es decir, más de 130.000.

El poco fruto obtenido de tantas horas de trabajo, lejos de ser desalentador, es muy prometedor: mi sondeo ha servido para descubrir todo un tesoro neológico. Hay muchísima neología en *Asturnews*, tanta que exigirá técnicas más avanzadas de filtrado, que habrá que aplicar, seguramente, en un momento previo al contraste de listas. Por fortuna, este reajuste —con el que arrancará la fase de prospección— podrá beneficiarse de la experiencia ganada en el proceso de prueba-error que ha supuesto el sondeo.

Al igual que los corpus y los programas diseñados, el listado completo de formas se ha adjuntado en formato digital a este trabajo. No he querido adjuntar ninguna de las listas secundarias, ya que muchos de los procesos de filtrado de ruido aplicados han sido masivos o meramente tentativos, por lo que podrían haberse producido pérdidas de información. Los resultados que indico en el siguiente apartado también han de tomarse con cautela, ya que proceden precisamente de listados provisionales.

4. Resultados

Como acabo de adelantar en el cierre del capítulo anterior, el volumen total de formas diferentes detectado en ambos corpus fue tan grande que hizo imposible un análisis completo de resultados.

Durante los procesos de prueba-error destinados a filtrar parcialmente los resultados, sin embargo, sí fue posible realizar ciertas observaciones que permitieron elaborar al menos una tentativa cualitativo-cuantitativa con datos bastante relevadores.

Su obtención se realizó mediante un borrado masivo de los siguientes elementos: todas las unidades con 1 ó más apariciones en cualquiera de los campos del DALLA (campos 3 y 4, como queda explicado en el apartado 3.3), unidades con patrones gramaticales en apariencia castellanos, unidades compuestas por caracteres mixtos (números y signos, además de letras).

Tras ello, con ayuda de Word y Excel, se ordenó el listado completo para que mostrara en primer lugar las palabras minúsculas de *Asturnews* ordenadas por frecuencia y, a continuación, las mayúsculas. Se tomaron después las palabras con mayor frecuencia de cada tipo de *caja* y se juntaron en un archivo nuevo, que se ordenó primero por orden alfabético (para eliminar mayúsculas innecesarias, como palabras comunes o nombres propios de persona) y luego por orden de frecuencia.

Finalmente, se seleccionaron las primeras 200 palabras de la lista y se realizaron varias agrupaciones por lemas (por ejemplo, todas las apariciones de *prósimu*, *prósimes*, *prósimos*… se sumaron en *prósimu*). Asimismo, en una revisión manual, se eliminaron falsos neologismos y se redujo el número de frecuencia de otros que realmente escondían formas de neología semántica (para detectarlos, se los sometió a una concordancia en contexto o KWIC).

Tras todos esos procesos, se obtuvo la siguiente lista ordenada (no indico los números de frecuencia ya que no son absolutos: al eliminar la parte mayoritaria de la lista, fue imposible sumar todas las formas de cada lema presente):

1. *prósimu*
2. *director*
3. web / Web
4. *cellebrar*
5. internet / Internet
6. rock
7. folk
8. recital
9. *hestóricu*
10. *alrodiu*
11. durante
12. *relación*
13. debate
14. *directu*
15. *cellebración*
16. Esteriores *(Asuntos …)*
17. eventu
18. dvd / DVD
19. varios
20. *administración*
21. jazz
22. Bruseles
23. difusión
24. ayudes *(… económiques)*

25. tv / TV
26. piesll<u>e</u>
27. respective
28. másimu
29. portavoz
30. collectivu
31. Executivu *(~gobiernu)*
32. bandina
33. radiotelevisión
34. sigún
35. reunión
36. pop
37. Secundaria *(Educación …)*
38. izquierdes (d')
39. Verdes *(~ecoloxistes)*
40. conceya<u>lía</u>
41. pesie
42. colaboración
43. garrapiellu
44. Administración
45. formaciones *(… polítiques)*
46. cantautor
47. acciones
48. blues
49. orixen
50. conceyal<u>a</u>
51. lema
52. UE (también XE)
53. Asturianía *(~asturianos mundo)*
54. reciente
55. respectivamente
56. <u>asinamesmo</u>
57. promocionar
58. asamblea
59. exposición
60. cualaquier
61. mediaos
62. coficialidá
63. acceder
64. sobremanera
65. tecnoloxíes
66. Xuníos (Estaos)
67. IES *(l')*
68. precisó *(~matizó)*
69. blog
70. conservatoriu
71. asoleyó *(~editó)*
72. vital
73. punk
74. residentes
75. comparecencia
76. competencies *(… autonómiques)*
77. software
78. xunidá
79. amues<u>a</u>
80. folclore
81. municipios
82. oficial<u>ización</u>
83. acoyerá *(~albergará)*
84. directiva
85. promocional
86. cual
87. collectivos
88. municipiu
89. temática *(sustantivo)*
90. CD
91. hip hop
92. roll
93. debatir
94. ociu
95. alpenes
96. cásique
97. cortometraxes
98. difundir
99. Rector
100. súmase *(ta a favor)*
101. directamente
102. dixitales
103. idega
104. itinerante
105. posibilistes
106. vigor (en)

Lo primero que llama la atención es la gran cantidad de falsos neologismos que, en realidad, son variantes ortográficas de las formas correctas (en concreto, son las formas indicadas totalmente en cursiva, como *prósimu*, *director*, *cellebrar*…, variantes de *próximu, direutor*, *celebrar…)*.

El fenómeno denota cierta inseguridad por parte de los usuarios, probablemente por las ambigüedades que todavía acompañan a las normas de la ALLA sobre el tratamiento de los grupos cultos. Así, mientras la mayoría de asturianos pronuncian *direZtor* y han aprendido a escribir *direC*tor (en castellano), la norma les exige que escriban *direUtor*, al estilo de la pronunciación occidental y más alejado gráficamente del castellano[63]. Sin embargo, en casos como *aCtuar* o *reCtor* la Academia ni siquiera plantea las opciones *aUtuar* y *reUtor*.

El tipo de desviación anterior, que podríamos llamar *cultista*, se ve complementado por otro, que podría decirse *hiperasturianista*[64], por el que se prefieren formas de aspecto más *legítimamente asturiano*. Ése es el caso de *prósimu*, *cellebrar*, *cualaquier*, *hestoria*…, frente a las formas indicadas como correctas o preferidas por la Academia: *próximu*, *celebrar*, *cualquier*, *historia*…

La abundancia de este tipo de formas nuevas es una señal de que aún queda mucho para que los procesos de alfabetización de los hablantes, por un lado, y de normativización, por otro, acaben de cuajar[65].

Otro fenómeno que indica la debilidad denominativa del asturiano es la abundancia de castellanismos del tipo de *debate*, *reunión*, *portavoz* o *difundir*, para los que la Academia propone *alderique*, *xuntanza*, *voceru* o *esparder*. Frente a la forma considerada tradicional (por ejemplo, *alderique*), muchos usuarios parecen preferir la forma más cultista (*debate*), bien por su parecido con el castellano, bien por ignorancia, bien por su mayor transparencia o bien por su semejanza con otras lenguas romances. En el ejemplo elegido, *alderique* presenta 197 apariciones absolutas en minúsculas, frente a las 244 de *debate*.

Otros neologismos, como el genial calco *asinamesmo*, o las formas, consideradas castellanizantes, *durante*, *municipios* y *varios* también serían muestra de la debilidad que aún caracteriza al asturiano. Se trata de palabras con una gran frecuencia relativa en castellano que, en cierto modo, parecen faltarle al escritor de asturiano. ¿Cómo decir *La conocí durante el verano*? ¿*Conocíla durante'l branu*? ¿*Conocíla en branu*? ¿*Conocíla esti branu*…? ¿Y *Varios municipios*? ¿*Dellos municipios*…? La preferencia por el castellano sería en este caso un probable índice de diglosia.

[63] Es la clásica tensión entre los conceptos sociolingüísticos de *Abbausprache* (lengua por elaboración) y *Abstandsprache* (lengua por distancia). En este caso, a falta de una distancia suficiente hacia la *lengua techo*, la diferencia se marcaría por elaboración, artificialmente.

[64] Me inspiro aquí en el concepto de *hiperenxebrismo*, utilizado en gallego para explicar precisamente el mismo fenómeno.

[65] Otro ejemplo muy interesante de la falta de consolidación de la normativa asturiana está en *curtiumetraxe*, forma correcta frente a la que figuran otras como *cortometraxe*, *curtiometraxe*, o *curtiumetrax*.

Pero no todo es señal de debilidad. Hay neologismos que indican que la lengua asturiana tiene capacidad para seguirse inventando, bien con recursos propios, más o menos inspirados (como *Asturianía* con mayúsculas, los deverbales *pieslle*, *amuesa = cierre*, *muestra*, los derivados *conceyala, conceyalía* y *oficialización* o el clásico *Bruseles*[66], mucho más natural para el oído asturiano que *Bruselas*…), bien con la adaptación natural de préstamos totalmente extranjeros y ajenos a la estructura fónica de la *llingua* (*blog*[67], *punk*, *software*…).

Otros neologismos indican la flexibilidad con que, siguiendo el ejemplo del castellano, las palabras asturianas van adquiriendo nuevos significados acordes con los nuevos tiempos y aún no recogidos por el DALLA. Es el caso de *Executivu*, *acoyer*, *verde*, *competencies*…

Por último, fuera de esta lista quería señalar un par de casos interesantes.

El primero es la abundancia de neologismos sintagmáticos formados por calco directo del castellano. Me refiero a formas como *dexar/poner en entredicho*, *entrar/tar en vigor*, *nel seno de*…, *a lo sumo*, *lo más granao*, *a lo bonzo*… y otros que sólo consiguen detectarse indirectamente.

El otro aspecto destacable es el de la riqueza formativa muchos prefijos y sufijos asturianos, como el semiculto *astur-*, empleado incluso con frecuencia para nombres comerciales[68]: *asturchale/asturtzale*, *asturcine*, *asturblog*, *asturfobia*, *asturlinux*, *asturmafia*… y el popular *-iegu*: *nacionaliegu*, *eonaviegu*, *españoliegu*, *internacionaliegu*, *rexonaliegu* e incluso *kosoviegu*.

Ambos ejemplos son una muestra del rico sistema de derivación del asturiano, al que la gramática oficial (Academia de la Llingua Asturiana: 2001) dedica un capítulo completo.

[66] Este aparente neologismo parece tener una base social muy extendida, fruto de la abundante emigración asturiana a Bélgica.
[67] Con menor frecuencia absoluta (21 apariciones frente a 67), aunque en aparente expansión, también circula la versión *blogue*, más ajustada a la prosodia asturiana.
[68] En el mismo sentido que *euro-*, *bio-* o *eco-* en otras lenguas.

5. Conclusiones

Ha llegado, por fin, el momento de las conclusiones, que dividiré en tres partes: una relativa a los objetivos planteados al inicio del trabajo, otra referente al trabajo en sí y, por último, una tercera consistente en un dictamen, a priori, de las grandes tendencias de la neología en asturiano.

Si regresamos a los objetivos, podremos ver que se han cumplido de forma positiva. En primer lugar, el objetivo general de conseguir una aproximación a la cuestión de la neología asturiana a partir de dos corpus (uno de uso real y el otro de carácter normativo) parece haberse alcanzado, quizá con unos resultados mucho más esperanzadores de lo previsto.

En efecto, creo que con mi sondeo he descubierto tal cantidad de posibles neologismos que va a hacer falta mucho tiempo, o bien unas estrategias muy afinadas, para poder analizarlos con profundidad. La primera aproximación, pues, ha sido reveladora: *le he visto las orejas al lobo*.

La experiencia de los errores cometidos y de los preanálisis realizados me obligará a plantear unos objetivos muy concretos para la fase de prospección, que permitan una explotación real de la enorme masa de *mineral* descubierta.

En cuanto a los objetivos secundarios, no me corresponde a mí determinar si se van a cumplir o no. Sí espero, no obstante, que este trabajo adquiera cierta repercusión, bien por el abrumador volumen de resultados obtenidos, bien por los materiales con que he trabajado. Sería muy satisfactorio saber que mi trabajo ha servido de inspiración, punto de partida o base para otras investigaciones. Como decía al principio de este sondeo, lo importante es que la rueda de la normalización se siga moviendo.

En lo que se refiere al trabajo, a las horas y el esfuerzo dedicado, creo que ha merecido la pena, por muy agotador que haya resultado. Decía en un punto anterior que mi intención era realizar un trabajo *a la escala del planificador particular*, es decir, dentro de unas posibilidades razonables. Bien, en este caso creo que sobrestimé mis fuerzas y, tal vez, subestimé los resultados que iba a obtener. Al menos ha quedado una experiencia pionera: se escribe mucho en asturiano, se escribe de forma muy innovadora con respecto a la *norma*, por lo que cualquier análisis medianamente ambicioso que se realice deberá afrontarse en equipo, so pena de sufrir grandes desvelos. Bromas aparte, creo que la enseñanza en este sentido debería estar en la importancia de la coordinación, de la colaboración desinteresada: si una sola persona es capaz de recopilar 2,4 millones de palabras y de extraer unos 130.000 candidatos a neologismos a partir únicamente de una fuente, ¿qué no podrán hacer *n* personas? Queda mucha prensa digital, mucha blogosfera, mucha *Uiquipedia* por analizar…

Por último, aunque no de forma menos importante, querría aventurar un pequeño diagnóstico del estado de salud del asturiano, a partir de los resultados parciales obtenidos con mi estudio.

En primer lugar, quiero señalar las fuertes tensiones que se detectan en el interior del lenguaje, causadas por una diglosia potentísima, por un déficit considerable en formación de los hablantes, por el peso de un castellano cada vez más fuerte y por la debilidad e inconsistencia interna que aún presenta el asturiano…, todos aspectos de difícil solución mientras no cambien las tornas políticas. El desarrollo de la Red colaborativa —aspecto en el que he insistido a o largo de todo el trabajo—, el impulso de debate interno y la voluntad de los hablantes, sin embargo, podría aportar cierta luz al último de los problemas: si hay inconsistencia es porque no hay suficiente debate, porque falta conocimiento. La solución, pues, podría estar más cerca de lo que pensamos.

Otro aspecto destacable es el de la influencia que ejerce el castellano, ineludible y no siempre negativa. Me explico: es negativa cuando supone un abandono o una degeneración de los recursos internos del asturiano, de su capacidad de renovación; es negativa cuando se ejerce desde situaciones de diglosia o de represión, directa o indirecta.

Ahora bien, también ejerce un efecto dinamizador muy positivo. Primero, *por acción*, porque abre la vía a cantidad de innovaciones cuyo coste de adaptación, al menos parcial, es relativamente bajo, dada la gran proximidad entre ambas lenguas (el castellano haría aquí de *hermano mayor* que va abriendo el camino a su *hermanito*). Pero también *por reacción*, ya que fuerza a los hablantes a buscar continuamente recursos propios, formas distintivas de marcar distancia que mantienen a la *llingua* en permanente evolución.

Un aspecto final sería el de conseguir que la *llingua* saliera de su rincón y pudiera sacar partido del enorme privilegio que constituye ser miembro del club de lenguas romances, siguiendo el ejemplo y los pasos de *grandes* como el castellano, el catalán o el portugués, y sirviendo de muestra e inspiración para otros *pequeños,* o simplemente *empequeñecidos*.

6. Bibliografía

Neología

- Cabré, T., Freixa, J. y Solé, E. (ed.) (2002): *Lèxic i neologia*, Observatori de Neologia (IULA), Universitat Pompeu Fabra, Barcelona.

- Gaudin, F. (2007): «Quelques mots sur la socioterminologie», en *Cahiers du Rifal (nº 26)*, Organisation internationale de la francophonie y Communauté française de Belgique, París y Bruselas.

- Observatori de Neologia (IULA) (2004a): *Llengua catalana i neologia*, Meteora, Barcelona.

- Observatori de Neologia (IULA) (2004b): *Metodologia del treball en neologia: criteris, materials i processos*, Observatori de Neologia (IULA), Universitat Pompeu Fabra, Barcelona.

- Termcat, Centre de Terminologia (2006): *Recerca terminològica. El dossier de normalització*, Eumo Editorial, Vic (Barcelona).

Neología y lengua asturiana

- Academia de la Llingua Asturiana (2001): *Gramática de la llingua asturiana*, Academia de la Llingua Asturiana, Uviéu.

- Academia de la Llingua Asturiana (2005a): *Constitución Española. Estatutu d'Autonomía del Principáu d'Asturies*, Academia de la Llingua Asturiana, Uviéu.

- Academia de la Llingua Asturiana (2005b): *Normes ortográfiques (6ª edición revisada)*, Academia de la Llingua Asturiana, Uviéu.

- Academia de la Llingua Asturiana (2006a): *Delles propuestes pa nomes de persona*, Academia de la Llingua Asturiana, Uviéu.

- Academia de la Llingua Asturiana (2006b): *La llingua na Alministración y otros documentos*, Academia de la Llingua Asturiana, Uviéu.

- Academia de la Llingua Asturiana (2007): *Abreviatures, rotulaciones y propuestes d'espresión y llocución*, Academia de la Llingua Asturiana, Uviéu.

- Baizán, E. (1989): «La normativización del léxicu na enseñanza académica. Nivel oral», en *Lletres Asturianes (nº 33)*, Academia de la Llingua Asturiana, Uviéu.

- CANO, A.M. (1993): «La elaboración del DALLA (Diccionariu de l'Academia de la Llingua Asturiana)», en *Actes du XXème Congrès International de Linguistique et Philologie Romanes (Zürich, 1991) (vol. IV)*, A. Francke Verlag, Tubinga y Basilea (Alemania y Suiza).

- CANO, A.M. (1995): «El Diccionariu de l'Academia de la Llingua Asturiana (DALLA)», en *La llingua asturiana / La langue asturienne / La lengua asturiana*, Academia de la Llingua Asturiana, Uviéu.

- CANO, A.M. (1996): «El Diccionariu de l'Academia de la Llingua Asturiana (DALLA)», en *Revista de Filología Románica (nº 13)*, Universidad Complutense, Madrid.

- CANO, A.M. (2002): «¿Cómo se fixo'l Diccionariu?», en *Informe sobre la llingua asturiana*, Academia de la Llingua Asturiana, Uviéu.

- CARCEDO, A. (2001): *Léxico disponible de Asturias*, Universidad de Turku, Turku (Finlandia).

- D'ANDRÉS, R. (2001-2003): *Cuestiones d'asturianu normativu (vol. I a III),* Publicaciones Ámbitu, Uviéu.

- D'ANDRÉS, R. (2007): «Incorrecciones y contravenciones llingüístiques na narrativa asturiana d'anguaño», en CAMPAL, X.LL. (coord.) (2007): *La emancipación de la lliteratura asturiana. Crónica y balance de la narrativa contemporánea*, Conseyería de Cultura, Comunicación Social y Turismu del Principáu d'Asturies, Uviéu.

- D'ANDRÉS, R. (2008): *La innovació neològica a l'asturià actual. Estat de la qüestió*, comunicación presentada en el *I Congrés Internacional de Neologia en les llengües romàniques, CINEO 08* (Barcelona, 7-10 de mayo de 2008), Observatori de Neologia (IULA), Universitat Pompeu Fabra.

- PORTA, X. (2004): «Llingua y ciencia n'Asturies», en *Lletres Asturianes (nº 87)*, Academia de la Llingua Asturiana, Uviéu.

- VIEJO, X. (coord.) (2008): *La ciencia n'otres pallabres*, Editorial Trabe, Uviéu.

- VILAREYO, X. (2001): «El modelu llingüísticu televisivu n'asturianu», en *Lletres Asturianes (nº 76)*, Academia de la Llingua Asturiana, Uviéu.

- VILAREYO, X. (2008): *Llingües vives y llingües muertes*, artículo publicado en http://www.xuristes.as, sitio web del Aconceyamientu de Xuristes pol Asturianu.

Neología e ingeniería lingüística

- Alonso, A., De Yzaguirre, Ll., Folguerà, R. y Tebé, C. (2002): «La mesura de la implantació terminològica: dades, variables i resultats», en IULA (2002): *Actes de la I Jornada sobre Terminologia i Serveis Linguistics (Barcelona, 18 de maig de 2001)*, Universitat Pompeu Fabra, Barcelona.

- De Yzaguirre, Ll., Matamala, A. y Cabré, M.T. (2000a): *El lematizador PALIC del IULA (UPF)*, comunicación presentada en el *XVIII Congreso de la AESLA* (Barcelona, 4-6 de mayo de 2000), Asociación Española de Lingüística Aplicada (AESLA).

- De Yzaguirre, Ll., Torner, S. y Matamala, A. (2000b): *El tratamiento automático de las ambigüedades segmentales del castellano*, comunicación presentada en el *XVIII Congreso de la AESLA* (Barcelona, 4-6 de mayo de 2000), Asociación Española de Lingüística Aplicada (AESLA).

- De Yzaguirre, Ll., Matamala, A.; Bach, C.; Castillo, N. y Ustrell, E. (2000c): *AMBILIC, el desambiguador lingüístico del Corpus del IULA*, comunicación presentada en el *XVIII Congreso de la AESLA* (Barcelona, 4-6 de mayo de 2000), Asociación Española de Lingüística Aplicada (AESLA).

- De Yzaguirre, Ll. (2001a): *La ingeniería lingüística de las lenguas minoritarias, un punto de vista catalán*, comunicación presentada en el *II Encuentro Internacional de los Multimedia y las Lenguas Minoritarias* (San Sebastián, 8-9 de noviembre de 2000), Gaia.

- De Yzaguirre, Ll., Tebé, C., Alonso, A., y Folguerà, R. (2001b): «El seguimiento de la implantación de términos vía Internet: estrategias de cálculo y control», en Correia, M. (2001): *Terminologia e Indústrias da Língua*, Fundação Gulbenkian, Lisboa.

- Janssen, M. (2005): *NeoTrack : Une analyseur de néologismes en ligne*, comunicación presentada en el *I Congrés Internacional de Neologia en les llengües romàniques, CINEO 08* (Barcelona, 7-10 de mayo de 2008), Observatori de Neologia (IULA), Universitat Pompeu Fabra.

- McEnery, T. y Wilson, A. (1996): *Corpus Linguistics. An Introduction*, Edimburgh University Press, Edimburgo.

- Nazar, R. (2007a): *Explotación estadística de corpus: análisis conceptual y clasificación de documentos*, comunicación presentada en el *Seminario IULATERM (Barcelona, 9 de febrero de 2007)*, Universitat Pompeu Fabra.

- NAZAR, R., VIVALDI, J. Y WANNER, L. (2007b): «Towards Quantitative Concept Analysis», en *Actas del XXIII Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) (Sevilla, 2007)*, Universidad de Sevilla, Sevilla.

- NAZAR, R. (2008): *Algunas técnicas de lingüística cuantitativa: el paquete Jaguar*, comunicación presentada en el *Seminario INFOLEX (Barcelona, 7 de marzo de 2008)*, Universitat Pompeu Fabra.

Anexos

Como complemento al trabajo realizado, se adjunta una carpeta zip con las siguientes subcarpetas:

1. **Corpus (3 archivos)**

   - **asturnews_DEFINITIVU.txt**: corpus de extracción completo
   - **dalla_DEFINITIVU.txt**: corpus lexicográfico de exclusión
   - **llista_DEFINITIVA.txt**: listado de unidades léxicas de los corpus

2. **Programación (11 archivos)**

   - **EstractorArticulu.pm**: módulo de marcaje para Asturnews
   - **ExtractorPalabra.pm**: módulo de marcaje para DALLA
   - **llimpia_articulos.pl**: programa de limpieza HTML en Asturnews
   - **llimpia_dalla.pl**: programa de limpieza HTML en Asturnews
   - **txt2Hash_DEFINITIVU.pl**: extractor de unidades léxicas
   - **URL_asturnews.txt**: URL de Asturnews generadas desde Perl
   - **URL_DALLA.txt**: URL de DALLA generadas desde Perl
   - **WGet_asturnews.txt**: comando de descarga en WGet de Asturnews
   - **WGet_dalla.txt**: comando de descarga en WGet de DALLA
   - **xenera_url_asturnews.pl**: generador de URL para Asturnews
   - **xenera_url_dalla.pl**: generador de URL para DALLA

3. **Proyecto (1 archivos)**

   - **sondeu_afcernuda**: texto completo del proyecto final de máster

- Notas -

1. *El contenido del corpus Asturnews (asturnews_DEFINITIVU.txt) es propiedad exclusiva de su autor (Asturnews, Próspero Morán López). Su uso para este proyecto de investigación se realiza según la licencia Asturnews. Para más información:*

   http://www.asturnews.com/popup_condiciones.php.

2. *El contenido del corpus DALLA (dalla_DEFINITIVU.txt) es propiedad exclusiva de la Academia de la Llingua Asturiana, según advertencia de copyright explícita en su sitio web: http://www.academiadelallingua.com.*

   Su uso se limita al presente trabajo de investigación. En ningún caso se realizarán copias o distribuciones a terceros sin el permiso expreso de la Academia de la Llingua Asturiana.

3. *Para la descarga gratuita de SCP: http://www.textworld.eu/scp/index.html.*

Milenta gracies

Llegáu'l momentu los agradecimientos, habrán perdoname los pocos llectores d'esti testu por dexar la llingua materna y pasar, sele seliquino, a la mio querida llingua güelerna.

Quede perclaro namás que nun hai diglosia nesta eleición —¡vállame Dios!—. Teo la rara suerte de ser monollingüe empederníu…, y, pa encima, na llingua l'Imperiu.

La mio intención, más bien, ye nun abandonar la llingua na que falo cola mayor parte los agradecíos y siguir cola mio xera de consiguir que los otros, por aburrición, lleguen a la deprender. Pocoñín a pocoñín, vais ver, acabaré aumentando'l númberu de falantes.

Darréu van, entós, los mios agradecimientos:

- *A los miembros del mio tribunal, los Dres. Rosa Estopà, José Enrique Gargallo y Lluís de Yzaguirre, pola so comprensión y paciencia nel momentu en que yeren más necesaries y por permitir que faiga la mio defensa nuna llingua proscrita.*

- *Al Dr. Lluís de Yzaguirre, en so calidá de direutor d'esti trabayu, pola so guía, pola so eterna paciencia col mio eternu pocu vagar y pol so compromisu heroicu y verdaderu pa con la mio llingua menor, pa con toes les llingües menores…*

- *A la Dra. Ana Cano y al Dr. Próspero Morán, pola so total disponibilidá pa ufrime'l frutu de tantu llabor ensin más esixencia que lu faiga valir pa qu'otros munchos sigan enanchándolu y faciéndolu espoxigar.*

- *A les Dres. Mª Teresa Cabré y Mercè Lorente, por abrime les puertes del IULA y descubrime les sos munches* ayalgues*, como Núria Bel, Jorge Vivaldi, Rogelio Nazar, Maarten Janssen y l'equipu completu de* iuleros*, del que, anque namás fuere por cuatro díes, fui arguyosamente ún más.*

- *Al Dr. Xosé Antón González Riaño, pola so visión de que'l futuru la llingua sedrá* online*, a Pili Fidalgo, por tar siempre ehí y al restu miembros y collaboradores de l'ALLA, pola so llucha incansable.*

- *A los Dres. Ramón d'Andrés y Xulio Viejo, por nun parar d'abrir víes nueves pa la llingua, por tener un espíritu*

- *A Rubén, a Mª Àngels y demás compañeros de* zulu*, por dame l'emburrión definitivu: ensin elli, taría tovía divagando…*

- *A mios collacios na llucha gayolera (Cuco, Esther, Manu, Lili y demás* ibicencos*), por comprendeme; a mios collacios cisastures y a mio familia, por sorprendese siempre.*

- *A Meri, insomne Meri (sábeslo perbién: agradecístimelo entós, agradecerételo siempre).*

- *Y, por supuestu…*

A los vieyos, por tenemos regalao tanto.