11-2020

Corretor ortográfico e corpus linguístico: matar dois coelhos com uma só cajadada

Matthew S. Stuckwisch
University of Tennessee at Chattanooga, matthew-stuckwisch@utc.edu

Línguas Minoritárias e Variação Linguística

Coordenação
Lurdes de Castro Moutinho
Rosa Lídia Coimbra
Alberto Gómez Bautista

universidade de aveiro
theoria poiesis praxis

FICHA TÉCNICA

TÍTULO
Línguas Minoritárias e Variação Linguística

COORDENADORES
Lurdes de Castro Moutinho, Rosa Lídia Coimbra, Alberto Gómez Bautista

EDITORA
UA Editora
Universidade de Aveiro
Serviços de Biblioteca, Informação Documental e Museologia

1.ª edição – novembro 2020

ISBN
978-972-789-656-1

DOI
10.34624/rj68-vz44

APOIOS

universidade de aveiro
dlc departamento de línguas e culturas
clic centro de línguas, literaturas e culturas

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Este trabalho é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projeto UIDB/04188/2020

Nota introdutória

No dia 6 de dezembro de 2019, realizou-se, no Centro de Línguas, Literaturas e Culturas (CLLC) da Universidade de Aveiro, Portugal, as *I Jornadas em Línguas Minoritárias*, tendo como comissão organizadora os editores do presente volume, Lurdes de Castro Moutinho (Professora Associada, CLLC, Universidade de Aveiro), Alberto Gómez Bautista (Professor Adjunto Convidado do ISCAL/ CLLC, Universidade de Aveiro) e Rosa Lídia Coimbra (Professora Auxiliar, DLC/CLLC, Universidade de Aveiro).

A comissão científica do evento integrou ainda os professores Helena Rebelo (CLLC/Universidade da Madeira), Maria Teresa Roberto (CLLC, Universidade de Aveiro), Maria Teresa Tedesco Vilardo Abreu (Universidade Estadual do Rio de Janeiro), Maria Victoria Navas (Universidade Complutense de Madrid), Vera Ferreira (CIDLeS; ELAR, SOAS University of London) e Xosé Luís Regueira (ILG, Universidade de Santiago de Compostela).

Com este evento, pretendeu-se aprofundar temáticas relacionadas com a investigação científica sobre línguas minoritárias, no âmbito da descrição, prescrição e normalização, com foco em contacto linguístico, influência da língua dominante, descrição gramatical, estudos fonéticos, estudos prosódicos e aspetos sociolinguísticos, tendo os trabalhos incluído conferências, comunicações orais e sessão de pósteres.

De entre as comunicações apresentadas no evento, após uma revisão científica por pares, foram selecionadas algumas delas, sendo a sua publicação da responsabilidade da Comissão Organizadora do evento.

Os editores

CORRETOR ORTOGRÁFICO E *CORPUS* LINGUÍSTICO: MATAR DOIS COELHOS COM UMA SÓ CAJADADA

Matthew Stephen Stuckwisch

CORRETOR ORTOGRÁFICO E *CORPUS* LINGUÍSTICO: MATAR DOIS COELHOS COM UMA SÓ CAJADADA

SPELL CHECKER AND LINGUISTIC CORPUS: KILLING TWO BIRDS WITH ONE STONE

Matthew Stephen Stuckwisch

(Department of Modern & Classical Languages & Literatures,
University of Tennessee at Chattanooga)

Resumo
O que faz falta para o estudo das línguas minoritarias é ainda muito. Uma das ferramentas mais importantes para o estudo de idiomas é o *corpus*. Embora seja bastante fácil hoje preparar um *corpus* básico, é bastante mais difícil criar um *corpus* etiquetado para a pesquisa de estruturas mais gerais, porque é preciso etiquetar as palavras nele introduzidas. Ao mesmo tempo, essas línguas também costumam estar num estado variável de normalização e estandarização, e elas e os seus falantes poderiam beneficiar de um corrector ortográfico. Proponho que, por mor de economia de recursos, é recomendável que quem quiser desenvolver um *corpus* também pense em fazer um corretor já que o trabalho para ambas as tarefas é muito parecido, senão quase idêntico. Este processo é aqui demostrado empregando o asturiano como exemplo.

Palavras-chave
Corretor ortográfico, *corpus* linguístico, asturiano, línguas minoritárias.

Abstract
There is much missing in the study of minority languages. One of the most important tools for the study of languages is the corpus. Although today it is easy to prepare a simple corpus, it is more difficult to create a tagged corpus because it is necessary to tag each of the words in it. At the same time, these languages tend to be in a variable state of normalization and standardization, and they and their speakers can benefit from a spell checker. I propose that, in the face of limited resources, it is best that a researcher intending to develop a corpus also consider making a spell checker, as the work for both is very similar, if not virtually identical. This process is demonstrated using Asturian as an example.

Keywords
Spell checker, linguistic corpus, Asturian, minority languages.

INTRODUÇÃO

A investigação linguística requer muitas ferramentas diferentes que, por vezes, não são fáceis de criar, porque obrigam tanto a habilidades informáticas quanto a conhecimentos linguísticos. Para as línguas minoritárias, os recursos costumam ser muito limitados, o que, frente aos recursos disponíveis para as línguas dominantes e quando for possível, implica combinar esforços. No presente trabalho, o enfoque será em duas ferramentas linguísticas fundamentais que podem e devem desenvolver-se juntas para reduzir o investimento necessário para ambas, especialmente quanto ao apoio informático. Seguidamente, com o fim de demonstrar a facilidade do processo, presentar-se-á um caso exemplar, e a importância deste método, usando a língua asturiana.

1. A IMPORTÂNCIA DOS CORRETORES ORTOGRÁFICOS

O propósito de um corretor ortográfico é simples. Ao processar uma sequência de palavras, deve detetar as palavras que foram escritas incorretamente e, se for possível, oferecer algumas recomendações para bem escrevê-las. É uma ferramenta que, para os falantes das principais línguas do mundo, é quase omnipresente no mundo informático, a ser incluída na maioria dos editores de texto e, mais recentemente, diretamente nos sistemas operativos, para ser usada em qualquer aplicação.

Mas a ferramenta é também importante no processo de estandardização de uma língua. Como já mencionado, os utilizadores de computadores, tablets ou telemóveis estão já muito acostumados a tê-la à sua disposição. Aliás, segundo várias organizações como a Digital Language Diversity Project (Berger et al., 2018), um corretor ortográfico é um dos primeiros passos para a habilitação de línguas minoritárias no mundo informático. Sem esse recurso tecnológico, é difícil imaginar o uso da língua dentro das tecnologias.

Mesmo entre línguas dominantes e semelhantes, o corretor ortográfico também pode empregar-se com o fim de evitar interferência negativa de outras línguas, enquanto promove a interferência positiva. Nas Astúrias, por exemplo, muita gente a nível popular fala *amestáu*, que é um dialeto que mistura o castelhano e o asturiano. Tal fenómeno é normal em qualquer região de contacto ou fronteiriça. A falta de escolarização na língua asturiana dificulta-lhes saber que palavras ou até estruturas gramaticais pertencem aos idiomas específicos. Com um corretor, ao introduzir uma palavra como *cerdo*, o utilizador poderá saber que não corresponde ao asturiano (sendo *gochu*), porque não seria reconhecida nem poderia oferecer nenhuma sugestão adequada. Do mesmo modo, se não souber a palavra no asturiano, mas sim no castelhano, pode comprová-la e, se for

reconhecida, beneficiará de uma interferência positiva confirmada. Assim, ao escrever *alumna,* ficará confirmada a validez da palavra e, ao escrever *gato,* terá, entre outras, a sugestão de *gatu*, lembrando ao utilizador que, no asturiano, os sustantivos masculinos singulares possuem a flexão *-u* e não *-o* como em castelhano.

Tabela 1. Diferentes versões de uma frase em idiomas ibéricos

língua	frase
asturiano	La xente que nun defende la llingua asturiana sedrán los responsables de la so destrución.
castelhano	La gente que no defiende la lengua asturiana serán los responsables de su destrucción
mirandês	La giente que nun defiende la lhéngua sturiana seran ls respunsables de la sue çtruçon.
português	A gente que não defende a língua asturiana serão os responsáveis da sua destrução
asturiano (com erros[a])	La xente que nun *defiende* la *lingua* asturiana sedrán los responsables de la *su* destrucción

[a] Os erros —por interferência castelhana— realçam-se em tipo itálico.

Como deve ser evidente no exemplo na Tabela 1, a língua asturiana é muito parecida com as línguas dos seus arredores. O castelhano, língua dominante, exerce uma influência preocu- pante sobre o asturiano e, por causa dele, não será infrequente encontrar os erros da Tabela 1 em textos escritos ou mesmo, algumas vezes, na comunicação oral dos falantes. Como já mencionado, um corretor ortográfico pode destacar ditas interferências e oferecer boas sugestões se se tratar de formas parecidas (na Tabela 1, deve-se escrever *defende*, *llingua* e *so*). A falta de sugestões, por outro lado, poderia indicar a inexistência de uma palavra cognata. Este benefício não é exclusivo do asturiano: para outras línguas como o escocês (que é semelhante linguisticamente à língua inglesa) ou o mirandês (com o português), será igualmente aplicável.

Mas as línguas que não sofrem desta interferência, por serem muito diferentes das línguas maioritárias onde se falam, também podem beneficiar de um corretor. Apesar de não compar- tilharem grande quantidade de palavras com as línguas maioritárias que muitas vezes também dominam os seus falantes, o corretor ajuda com o processo de estandarização e correção. Embora não seja um professor, o ato de confirmar, corrigir e sugerir modela o processo de aprendizagem escolar, até para pessoas mais velhas.

2. A IMPORTÂNCIA DO *CORPUS* ETIQUETADO

A outra ferramenta fundamental na investigação linguística é o *corpus*. Mas nem todos os *corpora* são iguais. Têm vários níveis de estrutura, sendo, na forma mais simples, um conjunto de palavras que permite a busca de palavras exatas, talvez com *wildcards* ou sequências simples. Quanto mais fixa a ortografia interior das palavras numa língua, mais fácil será pesquisar uma determinada forma (permitindo que, por exemplo, no mirandês *corr** coincida com *corre*, *corrimos*, *correrdes*). No entanto, para línguas que apresentam modificações internas, como alternâncias vocálicas, harmonia vocálica ou outros tipos de apofonia, será mais difícil. A pesquisa *corr**, que pretende encontrar todas as palavras que são formas de *correr*, por exemplo, não serviria para as formas do presente, pois este apresenta uma alternância vocálica *o-uo* (*cuorre*). Aliás, tanto no mirandês quanto no português, coincidiria com diversos substantivos relacionados, como *corrida*, mas também *correspundença*/*correspondência*, *corrupção* (port.) ou *corrícado* (mir.). Tais pesquisas podem destacar muitas palavras ou sequências não desejadas, ao mesmo tempo que falham em descobrir outras que sim se desejam.

Uma forma de melhorar as buscas é estabelecer um critério para captar só formas verbais, ou só palavras com alguma base particular. Assim, permitem-se pesquisas mais complexas e nítidas. Para línguas pouco ou nada flexionadas, como o inglês ou o chinês, é fácil construir um *corpus* com metainformações, por criar e consultar uma lista de metadados para cada palavra, porque o número de palavras é relativamente pequeno, dado cada lexema ter poucas formas. A palavra com mais variantes morfológicas para o inglês seria *be*.[1] Para cada palavra, portanto, podem ser especificados os seus metadados à mão.

Mas nas línguas mais flexionadas, como as línguas românicas, ou especialmente nas línguas aglutinativas, como muitas das línguas das Américas, se não é impossível, é suficientemente difícil para quase não valer o trabalho. Um verbo transitivo em português pode ter quase meia centena de formas, na flexão simples, e mais de mil contemplando os pronomes clíticos que, na correção ortográfica, se deverão contemplar. Não obstante, os padrões envolvidos costumam ser bastante fixos e previsíveis, o que permite a criação de um algoritmo que determine os mesmos metadados seguindo os padrões. Assim, uma palavra em português como *compravam* poderia ser processada como uma combinação dos elementos *compr-ava-m* (outras divisões seriam possíveis segundo modelo morfológico), com o primeiro elemento a indicar a palavra base *comprar*, o segundo a situar o tempo no pretérito imperfeito, e o terceiro *-m-* na terceira pessoa plural. Para qualquer flexão

[1] As formas de *be*, para além deste infinitivo, são: *am, art, is, are, was, wert, were, been, being*. Duas delas, *art* e *wert* já são arcaicas.

irregular, os metadados podem introduzir-se manualmente. Essa complexidade não seria talvez um grande problema, mas para calcular os metadados, seria preciso um informático para escrever um programa que, por sua vez, teria de consultar constantemente os linguistas para verificar se os pode calcular sem erro. Isso, claro, sem considerar os custos de o empregar.

3. DESENVOLVENDO AS DUAS FERRAMENTAS

Ambas as ferramentas podem implicar separadamente muito trabalho e obrigar a conhecimentos técnicos. Mas é evidente que podem ter diversos passos em comum, pontos de conexão onde é possível poupar quando se desenvolvem em conjunto.

Existem muitas bibliotecas diferentes para fins de correção ortográfica, mas o Hunspell (Németh, 2018), desenvolvido à base de MySpell (Hendricks, 2011), tem um alto nível de suporte em diversos sistemas operativos. Por esta razão, é recomendável o seu emprego quando os recursos informáticos (sobretudo na disponibilidade de peritos informáticos) são limitados e a compatibilidade é importante.

Os dados para Hunspell vêm em dois arquivos: um dicionário e uma listagem de afixos. Geralmente o dicionário contém uma lista de palavras numa forma base, com indicadores de flexão. A listagem de afixos, por seu lado, indica como modificar cada forma base para indicar pluralidade, género, etc.

O formato de Hunspell permite a inclusão de metadados para palavras e afixos para indicar a classe gramatical, mas é pouco utilizado.[2] O único dicionário conhecido neste formato que os contém é o húngaro — para o qual o Hunspell foi originalmente desenvolvido. Modificada minimamente a sintaxe para estes metadados, um exemplo de um afixo é o seguinte sufixo que serve para os adjectivos terminados em *-e*, que na forma feminina o troca por um *-a*.

```
SFX Ab Y 3
SFX Ab e a/Ll      e  gender:f   number:p
SFX Ab e es        e  gender:mf  number:p
SFX Ab e amente/Ll e                      category:adv
```

No dicionário, cada palavra leva indicações dos afixos admitidos para além dos metadados. A palavra *zoquete* tem a seguinte entrada:

```
zoquete/Ab  gender:mn  number:s  category:adj
```

[2] Por causa disso, o formato não é tão bem desenvolvido como poderia ser para os metadados relacionados. Dentro do código-fonte do presente projeto, pode-se ver com mais nitidez as ligeiras mudanças realizadas para facilitar a inclusão e processamento dos metadados necessários para o *corpus*.

Ela indica que a forma base é adjetivo singular semiambíguo quanto ao género (masculino ou neutro) e que admite as flexões atrás mencionadas, ou seja, *zoqueta* (feminino e singular), *zoquetes* (plural e ambíguo quanto ao género) e *zoquetamente* (advérbio).[3]

Para corrigir a ortografia, efetivamente o Hunspell tenta «desflexionar» a palavra, reduzindo progressivamente os afixos até ficar com uma palavra base. Com sucesso, a palavra é bem soletrada; e com fracasso, não, e um editor de texto pode assinalar ao utilizador, quer que a palavra não foi bem soletrada, se houver sugestões,[4] quer que não existe. Para criar um *corpus* etiquetado, é evidente que a maior parte do trabalho mais complexo já está realizada: ao «desflexionar» cada palavra num texto, somente é necessário recolher os metadados dos afixos, e associá-los à palavra base dentro do *corpus*.

4. O CASO DO ASTURIANO

Para demonstrar a possibilidade de facilmente combinar o trabalho para a criação do um corretor, usámos como caso exemplar a língua asturiana. O asturiano é uma língua do ramo asturleonês das línguas românicas e é falada na região setentrional de Espanha. É falado e entendido pela maioria dos habitantes das Astúrias, mas a alfabetização ainda não é universal entre os falantes. Em 2017, 62 % dos asturianos indicavam poder falá-lo, mas só 38 % indicavam uma capacidade de lê-lo e, ainda pior, apenas 25 % indicavam a capacidade de escrevê-lo (González Riaño et al., 2018, p. 147). Por isso, qualquer ferramenta que facilite a sua escrita — entre dar-lhes confiança ou proporcionar-lhes a mesma funcionalidade que têm para o castelhano — poderia fazer subir a proporção de asturianos que o escrevem.

Ao começar o projeto, o asturiano não contava com um corretor ortográfico adequado. Tinha um *corpus* grande mas básico e tinha uma vantagem importante no início deste projeto: a língua dispunha de um dicionário académico já publicado — O *Diccionariu de la Academia de la Llingua Asturiana* (2000). A Academia amavelmente ofereceu uma cópia digital do mesmo para facilitar este trabalho. Mesmo assim, ainda restava muito trabalho para preparar os dados, porque muitas entradas não vinham com toda a informação morfológica necessária para o projeto. Por exemplo, uma entrada para um substantivo ou adjetivo indicava quando uma palavra tinha duas ou

[3] É certo que, morfologicamente, a forma adverbial se forma à base da forma feminina pelo que, estritamente, deve haver uma sequência de dois sufixos. Aqui enfrentamos uma limitação de Hunspell, que somente permite dois sufixos no total. O afixo *LI* no exemplo adiciona o *ll*. Cada língua terá de determinar, no caso de serem possíveis mais do que três sufixos, como fundi-los no arquivo. Talvez no futuro a dita limitação seja eliminada, para os arquivos de afixos melhor refletirem a morfologia de cada língua.

[4] O método para gerir as sugestões é complexo e fora da avaliação. As sugestões de Hunspell são normalmente boas, mas, precisamente por ser uma ferramenta geral, as sugestões nunca vão ser tão boas como seriam se o corretor fosse feito exclusivamente para uma só língua (mas, para isso, seria preciso dedicar muito tempo e recursos ao seu desenvolvimento, algo que o presente trabalho tenciona evitar).

três variantes no singular — já que o plural é previsível a partir da forma singular — mas os verbos não indicavam o modelo de flexão, sobretudo se precisavam de ditongação. Apesar de muita desta informação estar presente nas *Normes Ortogràfiques* da Academia, ainda foi preciso introduzir manualmente os dados e provê-los quando não estavam disponíveis.

Neste sentido, o grupo SoftAstur prestou a sua colaboração. SoftAstur é uma organização dedicada à localização e tradução de *software* livre para o asturiano. Como acontece com muitos pesquisadores e voluntários na área de línguas minoritárias, há um nível linguístico alto, mas um nível tecnológico incompatível. Para facilitar a introdução de dados, em vez de pedir-lhes para escrever o arquivo Hunspell à mão, criou-se uma folha de cálculo na qual figuravam fórmulas que refletiam os afixos, a qual permitia uma pré-visualização das flexões (Figura 1).

Figura 1. Captura de ecrã da interface usada para introduzir metadados sobre substantivos

	A	B	C	D	E	F
1		FLEXONES	M	F	TIPU	IRREGULAR
2904	**ballena**	ballenes		X	reg.	
2905	**ballenatu**	ballenatos	X		reg.	
2906	**balleneru**	balleneros	X		reg.	
2907	**balleneru**	balleneros	X		reg.	
2908	**ballesta**	ballestes		X	reg.	
2909	**ballesteru**	ballestera ballesteros ballesteres	X	X	reg.	
2910	**ballestrinque**	ballestrinques	X		reg.	
2911	**balletón**	balletones	X		reg.	
2912	**ballicón**	ballicones	X		reg.	
2913	**ballicu**	ballicos	X		reg.	
2914	**ballique**	balliques	X		reg.	
2915	**balliqueru**	balliqueros	X		reg.	
2916	**ballocu**	ballocos	X		reg.	
2917	**balneariu**	balnearios	X		reg.	
2918	**balobru**	balobros	X		reg.	
2919	**balón**	balones	X		reg.	
2920	**balonazu**	balonazos	X		reg.	
2921	**baloncestista**	baloncestistes	X	X	inv.	
2922	**baloncestu**	baloncestos	X		reg.	
2923	**balonmano**		X		reg.	balonmanos
2924	**balsa**	balses		X	reg.	
2925	**balsamina**	balsamines		X	reg.	
2926	**bálsamu**	bálsamos	X		reg.	
2927	**balse**	balses	X		reg.	
2928	**balsera**	balseres		X	reg.	

Nota: A interface é simplesmente uma folha de cálculo (Google Sheets) que permitia a colaboração entre vários voluntários. Só precisavam marcar cada palavra como *M* (masculino) ou *F* (feminino), escolher o modelo de flexões (regular, invariável, ou irregular) e verificar se as flexões estavam bem formadas. Para as palavras irregulares, introduziam as flexões manualmente.

Para a maioria das palavras (as regulares), os colaboradores só tinham de marcar cada substantivo como masculino ou feminino. Se se tratasse de uma palavra irregular, as formas poderiam ser adicionadas individualmente. Desta forma, para além da criação inicial da folha, não

seriam precisos conhecimentos técnicos para a sua edição e poderiam trabalhar juntos muitos colaboradores.

Um *script* básico toma os dados da folha e produz um arquivo de dicionário Hunspell que se usa junto com os afixos predefinidos. O Hunspell vem já com um desflexionador que retorna palavras bases com os metadados correspondentes mas, para melhor integração no processamento, criou-se um analizador aparte que emprega os mesmos arquivos e dados.

5. RESULTADOS E EXEMPLO DE USO

Os resultados iniciais são já impressionantes. O pesquisador Matthew J. Burner está atualmente a desenvolver uma tese doutoral na Univeristy of Wisconsin–Madison sobre o emprego do neutro no asturiano (correspondência pessoal, 2019). Enquanto que os substantivos em asturiano normalmente são femininos ou masculinos, os adjetivos, na variedade central, possuem um terceiro género que indica abstração ou, quando aplicado a substantivos, incontabilidade ou materialidade. Curiosamente, o substantivo não muda e, portanto, para referir o que em português seria *a água fria*, diz-se *l'agua frío* (e não o feminino *fría*), para *o trabalho duro*, *el trabayu duro* (e não o masculino *duru*).

No momento, o investigador Burner realiza a sua pesquisa no *corpus* Eslema que é desenvolvido pela Universidá d'Uviéu numa colaboração dos departamentos de filologia e informática. Apesar do tamanho grande do *corpus* (mais de dez milhões de palavras), este não permite pesquisas segundo metadados para além de macrodados como género ou ano de publicação. Para Burner, há duas opções. A primeira é pensar em sequências de substantivos e adjetivos com os quais tem de realizar muitas pesquisas sem sucesso para encontrar alguma concordância. A segunda é pesquisar quer pelo substantivo, quer pelo adjetivo. Apesar de resultar em muitas concordâncias de texto, a plena maioria não serão exemplos da dita sequência porque um substantivo pode concordar no seu género inerente ou no neutro, e um adjetivo neutro pode não se referir a substantivos — resultando problemas parecidos com os descritos na secção sobre os *corpora*. Ambas as opcões têm algo em comum: fazem perder muito tempo.

Com a técnica descrita no presente trabalho, pôde fazer-se buscas em textos usando uma sequência muito precisa: substantivo singular seguido por um adjetivo cuja forma possivelmente é a do neutro, sem importância do lexema. Embora no seu caso exista uma possibilidade de recolher também adjetivos masculinos devido à ambiguidade em algumas palavras (por exemplo, *intelixente*, que é comum aos três géneros, ou *encantador*,[5] comum ao masculino e neutro), é possível limitar os

[5] Aos adjetivos em asturiano que terminam em *-dor* pode ou não acrescentar-se um *-o* no neutro. Ao feminino, como em português, acrescenta-se sempre um *-a*.

resultados a formas não ambíguas, ou seja, as terminadas em *-o*. Mas, mesmo com a dita ambiguidade na forma, se as concordâncias são visualizadas com contexto, normalmente é possível verificar se é um uso verdadeiro do neutro.

Para comprovar a técnica, fez-se a pesquisa *substantivo-singular* seguido por *adjetivo-possivelmente-neutro* com o texto íntegral de *Un vasu d'agua* de Ángela Carvajal (2017), um poemário de aproximadamente três mil palavras. Eis na Tabela 2 as concordâncias dentro do seu contexto.

Tabela 2. Resultados da pesquisa SUST-SING ADJ-NEUT$_{(possível)}$ no texto íntegral de *Um vasu d'agua*.

Texto anterior	Concordância	Texto posterior	Neutro?
sábanes felices Y la guasa	*de mio*	ma Se va el caimán se	x
Cantaba en cuantes me vía	*cola maleta*	nes manes Ónde andará'l caimán LA	x
al sentiles tiéntame nadar nesa	*agua doloroso*	descubrir la verdá que nun conozo	√
verdá que nun conozo paezme	*más urxente*	qu'escapar Y pégame otra vegada la	x
los sos peldaños la so	*madera claro*	serán d'otros los nuestros pasos Y	√
p'apertar nes manes pequeñes la	*rabia inocente*	VIDA Van yá dellos años nos	?
Ye güei rellumu mordiscu na	*fruta amargo*	de la muerte MARINA Llueve na	√
gris nun llagu d'agua verde	*y azul*	y qué guapes son les hores	x
Cerquina la espinera col so	*arume dulce*	arrodiando y embizcando'l corazón	?
y atendréme a los fechos	*rigor intelectual*	llámase esto pero namás ye descreimientu	?
estupor incompresible silenciu Conozo	*un probe*	dios ye un probe diablu al	x
Conozo un probe dios ye	*un probe*	diablu al que-y rezo con secretu	x
de la piel pero vago	*ente dos*	mundos que nun son DUERMES DIOS	x
Doite cuerda por pena y	*un poco*	pol ruíu que fai nel espaciu'l	x
páxines de pizarra escribe l'agua	*cola paciente*	claridá de quien yá too lo	x
de lo que foi piénsolo	*un poco*	per alto con una sonrisuca altiva	x
gota Ye bastante pa mi	*güei esti*	camín curtiu pel que llevo caminando	x
al fin Ye inestable'l mundu	*ye cierto*	y permanente la borrina ensimismao	x
CEMENTERIU INGLÉS Arbolón milenariu	*solombra verde*	tan grande como la vida Nun	?
misteriu la eternidá anda resignada	*y amable*	pente la yerba cuidando ensin que	x
y nun alcuentro Dame la	*mano too*	equí ye lo mesmo too equí	x
les plantes mentes escucho el	*movimientu musical*	númberu tres de Schubert unes cuantes	?
mentes escucho el movimientu musical	*númberu tres*	de Schubert unes cuantes veces de	x
de Schubert unes cuantes veces	*de siguío*	porque ye un puru vuelu Básta-y	x
flores que sostién L'agua verde	*y claro*	baxa burbusando y l'aire frescucu de	x
montesinos señaldosos consumíos	*pola so*	firida abrazaos a les nubes como	x
facer otra cosa l'agua verde	*y claro*	que baxa burbusando esa filigrana de	x
duelen Sabemos que cásique too	*dura poco*	que cásique nada ye necesario Mui	x
dura poco que cásique nada	*ye necesario*	Mui al nuestru pesar la nuestra	x
y entama a respirar esa	*borrina triste*	y suañador la nuestra vida VELA	?
una neña camina pel pasillu	*cola to*	lluz ente solombres y al llegar	x

Nota: Os extratos carecem de pontuação. Texto originalmente em maiúscula apresenta-se em versalete.
[a] Sabe-se que é neutro segundo a primeira palavra na concordância funciona como sustantivo no seu contexto e a segunda palavra funciona como adjetivo que (1) modifica o substantivo e (2) fica flexionado no neutro. Para indicar que não cumpre as condições em absoluto, emprega-se *x*, que sim as cumpre, usa-se √, e quando é impossível determinar por serem idênticas a forma neutra do adjetivo e a masculina/feminina, marca-se com *?*.

É fácil de notar, em alguns casos, que não se trata de sequências *substantivo-adjetivo*, devido a poderem ter formas idênticas alguns adjetivos e, por exemplo, verbos. A maioria dos resultados falsos é por conterem palavras frequentes não substantivos que podem, embora raramente, ser substantivos. Por exemplo, a palavra *y* quase sempre é conjunção coordenativa, mas também é definida pelo *Diccionariu* (2000) como substantivo feminino com a aceção de «Símbolu matemáticu [que representa un valor desconocíu]». Também se recolhem muitos exemplos de sequências cuja neutralidade é impossível de determinar, por conterem um adjetivo que não possui distinção entre as formas masculina e neutra. Por exemplo, ao dizer «borriña triste», a palavra *triste* não muda entre géneros. Apesar de que no dialeto central com que escreve Carvajal esperar-se-ia o neutro com algo incontável como *borriña* (*neblina,* em português) e ela usa-o de facto em frases como «ye cierto y permanente la borriña ensimismao»[6], é possível que ao falar de «esa borriña», ela a contemple como feminina por considerá-la contável: muitas palavras em asturiano podem ser empregadas quer como contáveis, quer como incontáveis.

No total, três das concordâncias são indubitavelmente casos de neutro, outras cinco poderiam ser, segundo a interpretação contextual, deixando vinte e um casos falsos. Tal resultado poderia parecer mau, mas deve lembrar-se que, se a pesquisa excluísse palavras que raramente são substantivos, como *de* ou *ye*, a proporção entre positivos e falsos seria muito aceitável.[7]

Na Tabela 3, vê-se a drástica redução quando a pesquisa se limita a adjetivos que têm forma própria no neutro: ocorrem os mesmos três positivos, mas já só oito casos falsos. Ao excluir palavras como *de* ou *un,* reduz-se a dois falsos. Com um etiquetador especializado, seria possível diminuir o número de falsos — em troca de tarefas técnicas custosas —, mas como nestes exemplos, o número de falsos não é ingovernável.

[6] A sequência *borriña ensimismao* é de facto exemplo de neutro mas não saiu nos resultados porque o verbo *ensimismar* não estava registado no *Diccionariu.* Eis aqui um exemplo de como o desenvolvimento do corretor poderia assinalar palavras que possivelmente precisam de registo. De igual forma, no futuro, o pesquisador poderia também acionar a base dos prefixos e sufixos quando a palavra base não aparece no dicionário.

[7] Para comparação, uma busca de só adjetivos definitivamente neutros (terminados em *-o,* sem consideração do seu possível uso adverbial) encontra 33 exemplos frente a 143 para uma de adjetivos possivelmente neutros.

Tabela 3. Resultados da pesquisa SUST-SING ADJ-NEUT$_{(indudável)}$ no texto íntegral de *Um vasu d'agua*.

Texto anterior	Concordância	Texto posterior	Neutro?
al sentiles tiéntame nadar nesa	*agua doloroso*	descubrir la verdá que nun conozo	√
los sos peldaños la so	*madera claro*	serán d'otros los nuestros pasos Y	√
Ye güei rellumu mordiscu na	*fruta amargo*	de la muerte MARINA Llueve na	√
Doite cuerda por pena y	*un poco*	pol ruíu que fai nel espaciu'l	x
de lo que foi piénsolo	*un poco*	per alto con una sonrisuca altiva	x
al fin Ye inestable'l mundu	*ye cierto*	y permanente la borrina ensimismao	x
y nun alcuentro Dame la	*mano too*	equí ye lo mesmo too equí	x
de Schubert unes cuantes veces	*de siguío*	porque ye un puru vuelu Básta-y	x
flores que sostién L'agua verde	*y claro*	baxa burbusando y l'aire frescucu de	x
facer otra cosa l'agua verde	*y claro*	que baxa burbusando esa filigrana de	x
duelen Sabemos que cásique too	*dura poco*	que cásique nada ye necesario Mui	x

Nota: Os extractos carecem de pontuação. Texto originalmente em maiúscula apresenta-se em versalete.
[a] A neutralidade do adjetivo é confirmada só com o facto de modificar a palavra anterior em caso de esta ser uma palavra que funciona como substantivo. Emprega-se √ quando é exemplo, e x quando não é.

As duas pesquisas não podem recolher todos os exemplos do neutro no texto, devido ao frequente uso de hipérbato na poesia, mas demonstram a capacidade de realizá-las de forma mais nítida que com um *corpus* básico e, mais, realizadas à base de um simples corretor ortográfico. Burner pode criar outras sequências para tentar captar outros exemplos do uso do neutro ou expandir as mesmas para fontes maiores de texto mais amplas e extendidas. Outros investigadores poderiam usar o pesquisador no *corpus* para, por exemplo, encontrar concordâncias de uma reduplicação de pronomes oblíquos independentemente de serem pronomes enclíticos ou proclíticos (mas assegurando que o pronome concorda com o substantivo correspondente). Ou, mais simplesmente, poderiam realizar buscas para palavras em todas as suas flexões.

Fora do mundo académico, os falantes não investigadores do asturiano podem desfrutar do corretor ortográfico dado as suas experiências como utilizadores de computador não serem nada diferentes das com o castelhano. Na realidade, vários autores asturianos já o estão a empregr (Xandru Martino Ruz, Xesús González Rato, correspondências pessoais, 2019), mediante uma versão para LibreOffice. Aliás, quando encontram uma palavra não esteja registada no dicionário, tomam nota, quer para corrigir algum erro nos arquivos, quer para ela ser incluída em próximas edições do dicionário — um não esperado benefício do corretor que pode também facilitar a produção de um dicionário para as línguas que não contam com um dicionário comprensivo. Sem dúvida que algumas das palavras descobertas desta forma serão introduzidas na próxima edição do *Diccionariu*.

6. TRABALHO FUTURO

Espera-se que futuras investigações sejam realizadas com os desenvolvedores do Eslema para permitir o seu emprego neste *corpus*: eles têm como aspiração, desde 2008, a capacitação de pesquisas mais detalhadas como as realizadas no presente trabalho. Para facilitar as ditas pesquisas, desenhar-se-á também uma sintaxe de pesquisa que qualquer investigador possa aproveitar, já que, de momento, as pesquisas têm de ser programadas individualmente com código bastante complexo, à espera de desenvolver ou adoptar uma linguagem de domínio específico.[8] Enquanto isso, as ferramentas de suporte criadas serão modificadas para serem mais generalizadas, bem documentadas, e distribuídas para facilitar trabalhos parecidos noutras comunidades linguísticas minoritárias (Stuckwisch, 2020).

CONCLUSÃO

Em resumo, atualmente a maioria dos dicionários em formato Hunspell — quer de línguas minoritárias quer de línguas dominantes —, carecem de informação linguística do tipo que facilitaria a criação de *corpora*[9]. Muitas línguas minoritárias nem têm corretor ortográfico, portanto, ao começar a criar um, é importante (e não difícil) combinar o trabalho para fazer um *corpus* etiquetado, visto que não requer muito trabalho adicional. Por exemplo, e sobretudo dada a sua semelhança com o asturiano, o mirandês e o aragonês poderiam quase copiar o modelo apresentado neste trabalho para aproveitar dos numerosos textos literários já escritos. Se para uma língua não existe nenhum dicionário, a poupança pode ser ainda maior: criando simultaneamente dicionário, corretor e *corpus*.

Apesar de que o processo pode precisar de um pouco de trabalho informático, conforme a experiência com o asturiano, o esforço do informático pode ficar muito reduzido enquanto a maioria do trabalho pode ser feito por pessoas que não possuam muitos conhecimentos informáticos e dentro de aplicações que já sabem utilizar. Assim, a expansão das ferramentas disponíveis tanto para os investigadores como para os falantes é possível numa só cajadada.

[8] O autor conhece a CQL (Corpus Query Language, ou linguagem para consultas de *corpus*) desenhada por SketchEngine (Lexical Computing, s.d.), que já integra muitas das características desejadas, pelo que uma opção seria extendê-la para as outras descritas também neste trabalho.

[9] Apesar de não existir um único sítio onde estejam todos os arquivos desenvolvidos para Hunspell, Titus Wormer (2020) inclui quase cem no seu repositório. Só o do húngaro fornece os metadados precisos para a criação de um *corpus* etiquetado.

REFERÊNCIAS BIBLIOGRÁFICAS

Academia de la Llingua Asturiana. (2000). *Diccionariu de la Academia de la Llingua Asturiana*. Uviéu: Academia de la Llingua Asturiana.

Academia de la Llingua Asturiana. (2014). *Normes Ortográfiques*. Uviéu: Academia de la Llingua Asturiana.

Berger, K. C., Gurrutxaga Hernaiz, A., Baroni, P., Hicks, D., Kruse, E., Quochi, V., Russo, I., Salonen, T., Sarhimaa, A. & Soria, C. (2018). *The LDP Digital Language Survival Kit*. Retirado de http://www.dldp.eu/sites/default/files/documents/DLDP_Digital-Language-Survival-Kit.pdf

Carvajal, A. (2017). *Un vasu d'agua*. Uviéu: Saltadera.

González Riaño, X. A., Fernández Costales, A., & Avello Rodríguez, R. (2018). Marcu llegal de la llingua asturiana y aspeutos sociollingüísticos y socioeducativos n'Asturies. Em *Informe sobre la llingua asturiana* (2.ª ed., pp. 139–181). Uviéu: Academia de la Llingua Asturiana.

Hendricks, K. (2011). *Myspell*. https://code.google.com/archive/a/apache-extras.org/p/ooo-myspell/

Lexical Computing (s.d.). *SketchEngine*. Retirado de http://www.sketchengine.eu

Neira Álvarez, X., Viejo Fernández, X., Ferández Lorences, T., Suarí Colomer, R., Fernández Rubiera, F., (s.d.). *Eslema*. Retirado de http://eslema.uniovi.es/corpus/busqueda.php

Németh, L. (2018). *Hunspell* (1.7.0). Retirado de http://hunspell.github.io

Stuckwisch, M. S. (2020). *HunspellTagger* (0.1). Retirado de http://www.github.com/alabamenhu/hunspelltagger/

Wormer, T. (2020). *Dictionaries*. Retirado de https://github.com/wooorm/dictionaries/