Dictionary Alignment by Rewrite-based Entry Translation

Alberto Simões[1] and Xavier Gómez Guinovart[2]

1 **Centro de Estudos Humanísticos, Universidade do Minho**
Campus de Gualtar, Braga, Portugal
ambs@ilch.uminho.pt

2 **Galician Language Technology and Applications (TALG Group)**
Universidade de Vigo, Galiza, Spain
xgg@uvigo.es

Abstract

In this document we describe the process of aligning two standard monolingual dictionaries: a Portuguese language dictionary and a Galician synonym dictionary. The main goal of the project is to provide an online dictionary that can show, in parallel, definitions and synonyms in Portuguese and Galician for a specific word, written in Portuguese or Galician.

These two languages are very close to each other, and that is the main reason we expect this idea to be viable. The main drawback is the lack of a good and free translation dictionary between these two languages, namely, a dictionary that can cover lexicons with more than one hundred thousand different words.

To solve this issue we defined a translation function, based on substitutions, that is able to achieve an F_1 score of 0.88 on a manually verified dictionary of nine thousand words. Using this same translation function to align a Portuguese–Galician dictionary we obtained almost 50% of the dictionary lexicon (more than eighty thousand words) alignment.

1998 ACM Subject Classification I.2.7 Natural Language Processing

Keywords and phrases translation, rewrite system, dictionary, dictionary alignment

Digital Object Identifier 10.4230/OASIcs.SLATE.2013.1

1 Introduction

Dicionário-Aberto[1] [10] resulted from the transcription, validation, and annotation of a printed dictionary for the Portuguese language, compiled by Cândido de Figueiredo, and published in 1913. It was transcribed as a Gutenberg Project book, but the main goal for this task was the use of this dictionary to bootstrap an XML-encoded dictionary that could be enriched and expanded by the community, and that can be used in Natural Language Processing (NLP) tasks. The document was subject to different steps on semantic annotation and orthography modernization [8], and is currently being used for the extraction of different NLP resources [11, 9]. It is also available in a web site for online querying, both as a standard form, and as a RESTless server.

In the context of another project [5], a synonym dictionary for the Galician language [7] (check also Guinovart and Simões, this proceedings) was converted from Microsoft Word files to a semantic-rich XML file. This dictionary was also corrected, widened in lexical extension, and modernized, taking into account the current norms for the Galician language. At the

[1] Available at http://dicionario-aberto.net/

2nd Symposium on Languages, Applications and Technologies (SLATE'13).
Editors: J. P. Leal, R. Rocha, A. Simões; pp. 1–11

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

moment the result of this work is not available to download because it is not finished yet, but the dictionary contents will soon be available in a web site for online querying.

Given the proximity of the two languages we decided that it would be interesting to show Galician definitions together with the Portuguese definitions. This would be useful for language researchers, but it can also be used to enrich a common thesaurus.

The main problem to bring this project to life is the alignment task: how to make entries from both dictionaries correspond to each other.

This task could be easy to execute if there was a bilingual dictionary. But there are few bilingual dictionaries between Portuguese and Galician, and the ones available are too small to allow the alignment of dictionaries with more than a hundred thousand entries.

The main contribution of this article is the test of the following hypothesis:

> *Given the proximity between the two languages, would it be possible to transform a Portuguese word in the dictionary into a Galician word just by applying a set of rewrite rules?*

The following section describes the translation function, what rewrite rules are used, and the order in which they should be applied. Section 3 describes two evaluation processes: the first one using a small Portuguese–Galician bilingual dictionary, that was hand-curated; the second one, using a bigger dictionary obtained by dictionary triangulation. In Section 4 the dictionary alignment process is performed, and the results discussed. Finally, we conclude with some final remarks.

2 Translation Function

The translation function that, given a set of valid Galician words (L_{gl}) and a Portuguese word (w_{pt}), returns a single[2] Galician word, will be denoted by $\mathcal{T}(L_{gl}, w_{pt})$, and is defined as a set of substitutions that rewrite a Portuguese word into a Galician translation.

Table 1 summarizes the performed substitutions. First, the Portuguese word is tried as a Galician word without any modification. If it is does not exist in the target lexicon, the substitutions are performed. The substitutions are ordered from more general substitutions to more specific ones (this was done manually, both from the authors knowledge of the two languages, and querying the dictionary to confirm the number of cases for each substitution). In some cases, less general rules needed to be performed first than more general ones, because of their interdependence. For example, the substitution from *-ção* > *ción* depends on the existence of the *ç* character that might be substituted by the *z* character, if the more general substitution *ç* > *z* is applied first. If they get applied in the wrong order the second substitution will not take place (as no *ç* character will be found), decreasing the number of correctly translated words.

Notice that some substitutions have two possible targets. In these cases, both possible words are maintained, and consequent substitutions will be applied to all words. That is why there are some substitutions that include the string being substituted as the substitution result: *im-* > *im-, inm-*. This rule will force that all Portuguese words with the prefix *im-* will be rewritten into two possible translations: the original one, and another one where *im-* was substituted by *inm-*.

[2] As explained below, the function generates, internally, a set of possible translations, but only one is returned.

At the end of the substitution process each possible translation is checked against the Galician lexicon (L_{gl}), and the first one that exists is returned (this one, w_{gl}, is the translation of w_{pt} using the translation function). This means that, internally, the translation function is over-generating words (both correct words and non-existent words), given that they can be filtered before returning, using the target lexicon.

To exemplify the rewrite rules, starting with the word *impassível*, we can derive:

$$\begin{array}{lll} \textit{impassível} & >_A & \textit{impasível} \\ & >_M & \textit{impasíbel, impasible} \\ & >_{AI} & \textit{impasíbel, impasible, inmpasíbel, inmpasible} \end{array}$$

The words from this list of generated words are then searched in the Galician lexicon and the first one that exists is returned as correct: *impasíbel*.

3 Evaluation

To evaluate the substitutions we performed two different runs, using two different dictionaries. The first one uses a small translation dictionary from Galician to Portuguese that was hand-curated. The second experiment was performed on a Galician–Portuguese dictionary obtained by triangulation using different pivot languages. The following sections explain the used metrics, detail the origin of these dictionaries, and present and discuss the obtained results.

3.1 Evaluation Metrics

The two evaluations were performed using hypothesis testing. Table 2 presents the Type I and Type II error matrix. Cell counts are computed as follows:

(TP) True Positives – a Portuguese word is correctly transformed by the translation function into one of the possible corresponding translations;

(FP) False Positives – the proposed translation for the Portuguese word is not the correct one, but is listed in the Galician lexicon (it is present in the dictionary as a translation for some other word);

(TN) True Negative – the proposed translation for the Portuguese word is not listed in the Galician lexicon, but is a correct translation. This can never happen because if the translation is correct, then it exists in the gold standard, and therefore, it will necessarily exist in the Galician lexicon (as it is computed from the gold standard). Thus, it is impossible to have such a word: $TN = 0$.

(FN) False Negative – whenever the proposed translation word does not exist in the Galician lexicon, and is not a correct translation. This happens every time the translation is not in the Galician lexicon (as it is computed from the translation pairs).

To evaluate the proposed substitutions we computed the usual metrics: accuracy, precision, recall and F_1 measure, using the standard formulae.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Table 1 List of the translation function substitutions, by application order.

Identifier	Substitution	Examples
ID		*mesmo* > *mesmo, normativa* > *normativa*
A	`ss > s`	*passo* > *paso*
B	`j > x`	*sujeito* > *suxeito, injectar* > *inxectar*
C	`-ção > -ción,-zón`	*adivinhação* > *adiviñación, coração* > *corazón*
D	`ç > z`	*laço* > *lazo, carroça* > *carroza*
E	`nh > ñ`	*unha* > *uña*
F	`-dizer > -dicir`	*contradizer* > *contradicir, desdizer* > *desdicir*
G	`z ([eiéíêî]) > c`	*bronze* > *bronce*
H	`lh > ll`	*alho* > *allo*
I	`vr > br`	*livro* > *libro*
J	`-agem > -axe`	*arbitragem* > *arbitraxe*
K	`g ([eiéíêî]) > x`	*faringe* > *farinxe, agência* > *axencia*
L	`-ável > -ábel,-able`	*amável* > *amable, amábel*
M	`-ível > -íbel,-ible`	*possível* > *posible, posíbel*
N	`-velmente > belmente,-blemente`	*previsivelmente* > *previsibelmente, previsiblemente*
O	`-eio > -eo`	*alheio* > *alleo*
P	`-ância > -ancia`	*abundância* > *abundancia, alternância* > *alternancia*
Q	`-ência > -encia`	*abstinência* > *abstinencia, agência* > *axencia*
R	`-aria > -ería,-aría`	*livraria* > *librería, libraría; tesouraria* > *tesourería, tesouraría*
S	`-ário > -ario`	*operário* > *operario, vestiário* > *vestiario*
T	`-óri[oa] > -ori[oa]`	*absolutório* > *absolutorio, aleatória* > *aleatoria*
U	`-são > -sión,-són`	*ilusão* > *ilusión, brasão* > *brasón*
V	`-rão > -rón,-rán`	*padrão* > *padrón, alcorão* > *alcorán*
W	`-mão > -món,-mán`	*limão* > *limón, caimão* > *caimán*
X	`-ião > ión,-ián`	*ancião* > *anción, anfitrião* > *anfitrión*
Y	`-ício > -icio`	*edifício* > *edificio*
Z	`-óide > -oide`	*asteróide* > *asteroide*
AA	`-ídio > -idio`	*presídio* > *presidio*
AB	`-ânico > -ánico`	*mecânico* > *mecánico*
AC	`-édia > -edia`	*comédia* > *comedia*
AD	`-cimento > -cemento`	*reconhecimento* > *recoñecemento* (always as suffix, not as a word)
AE	`-m > -n`	*além* > *alén*
AF	`-crever > -cribir`	*escrever* > *escribir, inscrever* > *inscribir*
AG	`-u > -u,-o`	*mau* > *mao, museu* > *museo, ateu* > *ateo*
AH	`-var > -bar`	*reprovar* > *reprobar* (when *-var* is kept, full word matches the PT word)
AI	`im- > im-,inm-`	*imortalidade* > *inmortalidade, improvável* > *improbábel*
AJ	`qua- > cua-,ca-`	*quanticamente* > *cuanticamente, quadro* > *cadro*
AK	`qua > cua`	*adequado* > *adecuado*
AL	`-xão > -xón,-xión`	*inflexão* > *inflexión, paixão* > *paixón*
AM	`rv > rv,rb`	*preservação* > *preservación, estorvar* > *estorbar*
AN	`-iver > -ivir`	*conviver* > *convivir, sobreviver* > *sobrevivir*

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

For better understanding of these measures, the evaluation tables presented in the next sections include two additional columns: one with the number of correct translations; and the other one with the number of additional correct translations generated by the application of that substitution.

3.2 Evaluation 1: Gold Standard

The rules were defined with a gold standard dictionary that was used to evaluate the substitutions relevancy, and the better sequence to use. For that purpose we downloaded a Portuguese–Galician translation dictionary from the Apertium project [4]. All multi-word sequences were removed, and a spell checker was used in the Portuguese portion of the dictionary to detect words written in the Brazil orthography (that were manually rewritten to the European Portuguese orthography), words that were written according to the Orthographic Agreement of 1990 (the dictionary to align uses orthography before 1990), and some other wrong words were also fixed.

After this cleaning process, the dictionary counts 9 224 pairs. Note that each pair maps a Portuguese word to a set of possible Galician translations. Table 3 presents the results. Each line refers to a different run, adding a new rule to the rule set. The first line, labeled as ID, corresponds to the first run, without any substitution. Looking into the accuracy for that line, one can see that 58% of the Portuguese words in the dictionary do not need translation, as they are shared across languages. In the second line the substitution A is activated ($ss > s$), leading to more 163 correct translations. For the third run, and before substitution B is ran, the system performs the substitution A. This means that each row includes the previous substitutions, and this explains the relevance of the delta column, which shows the number of accepted translations that each substitution generates[3]

There are some rules with a small delta, like rule Z. Nevertheless, the suffix *-oide* is specific of technical terms. The small dictionary used for this specific evaluation does not cover technical terms (with few exceptions), and we expect the rule to be more productive with a bigger dictionary.

At the end of the experiment we were able to keep the precision above 99.5% (higher than the obtained without any substitution) and a recall of 79.5% (compared with the 58.6% obtained without substitutions) resulting in a slight good F_1 measure.

[3] In fact, not exactly, as words might need more than one substitution to be correct.

Table 2 Hypothesis Type I and type II error matrix.

$\mathcal{T}(L_{gl}, w_{pt}) = w_{gl}$	Correct	Incorrect
w_{gl} is a Galician word	TP	FP
w_{gl} is not a Galician word	TN	FN

Table 3 Given 9 226 pairs mapping Portuguese words to a set of possible Galician words, the table presents precision, recall and F_1 measure; accuracy, total of correct words, and delta of correct words from last run. Note that substitutions are cumulative (meaning that when substitution B is performed, substitution A was performed before).

Subst. Id.	Precision	Recall	F_1	Accuracy	Correct	Δ
ID	0.9954	0.5859	0.7376	0.5843	5390	5390
A	0.9952	0.6038	0.7516	0.6020	5553	163
B	0.9951	0.6158	0.7608	0.6139	5663	110
C	0.9952	0.6567	0.7912	0.6546	6038	375
D	0.9951	0.6687	0.7999	0.6665	6148	110
E	0.9952	0.6782	0.8066	0.6760	6235	87
F	0.9952	0.6786	0.8070	0.6764	6239	4
G	0.9953	0.6838	0.8107	0.6816	6287	48
H	0.9953	0.6927	0.8169	0.6905	6369	82
I	0.9953	0.6934	0.8174	0.6911	6375	6
J	0.9953	0.6964	0.8195	0.6942	6403	28
K	0.9955	0.7210	0.8363	0.7187	6629	226
L	0.9955	0.7256	0.8394	0.7232	6671	42
M	0.9955	0.7284	0.8413	0.7260	6697	26
N	0.9957	0.7482	0.8544	0.7458	6879	182
O	0.9957	0.7496	0.8553	0.7472	6892	13
P	0.9957	0.7515	0.8565	0.7490	6909	17
Q	0.9957	0.7588	0.8612	0.7563	6976	67
R	0.9957	0.7602	0.8621	0.7577	6989	13
S	0.9958	0.7680	0.8672	0.7655	7061	72
T	0.9958	0.7703	0.8686	0.7678	7082	21
U	0.9958	0.7772	0.8731	0.7747	7146	64
V	0.9958	0.7780	0.8735	0.7755	7153	7
W	0.9958	0.7783	0.8737	0.7758	7156	3
X	0.9958	0.7796	0.8746	0.7771	7168	12
Y	0.9958	0.7806	0.8752	0.7781	7177	9
Z	0.9958	0.7807	0.8753	0.7782	7178	1
AA	0.9958	0.7813	0.8756	0.7787	7183	5
AB	0.9958	0.7818	0.8759	0.7793	7188	5
AC	0.9958	0.7822	0.8762	0.7797	7192	4
AD	0.9959	0.7836	0.8770	0.7810	7204	12
AE	0.9959	0.7855	0.8783	0.7830	7222	18
AF	0.9959	0.7863	0.8787	0.7837	7229	7
AG	0.9957	0.7876	0.8795	0.7849	7240	11
AH	0.9957	0.7882	0.8799	0.7856	7246	6
AI	0.9958	0.7903	0.8812	0.7876	7265	19
AJ	0.9956	0.7928	0.8827	0.7900	7287	22
AK	0.9956	0.7940	0.8834	0.7912	7298	11
AL	0.9956	0.7947	0.8839	0.7920	7305	7
AM	0.9956	0.7951	0.8842	0.7924	7309	4
AN	0.9956	0.7955	0.8844	0.7927	7312	3

3.3 Evaluation 2: Triangulated Dictionary

The dictionary used in the previous section is not a large dictionary. When trying to evaluate the translation algorithm in a bigger bilingual dictionary we hit a wall: the scarcity of free Portuguese–Galician dictionaries.

To solve this issue we performed triangulation with different dictionaries:

- Using the Portuguese–Spanish (12 340 pairs) and the Spanish–Galician (7 581 pairs) bilingual dictionaries from the Apertium translation software, resulting in a Portuguese–Galician bilingual dictionary with 5 045 pairs;
- Using the Portuguese–Spanish (12 340 pairs) and the Spanish–English (24 912 pairs) bilingual dictionaries from the Apertium translation software, and an English–Galician (17 626 pairs) bilingual dictionary from the CLUVI project [6], resulting in a Portuguese–Galician bilingual dictionary with 6 644 pairs;
- Using the Portuguese–English (14 600 pairs) from a merchandising application offered years ago by a beverages make, and the English–Galician (17 626 pairs) bilingual dictionary from CLUVI project, resulting in a Portuguese–Galician bilingual dictionary with 8 589 pairs.

These three dictionaries obtained, and the original Portuguese–Galician dictionary used in the previous section, were added together, resulting in a 14 492 pairs bilingual dictionary (5 268 more pairs than the original dictionary).

Before presenting the results, a brief explanation of how the triangulation process was performed, and how the dictionaries were merged together is in order:

- **Triangulation:** Each one of the dictionaries used in any of the triangulation processes contains lists of pairs, mapping words from the source language to a list of words in the target language. Therefore, the process needs two source dictionaries $\mathcal{D}_1 : L_S \mapsto \mathcal{P}(L_I)$ and $\mathcal{D}_2 : L_I \mapsto \mathcal{P}(L_T)$. For each word in the source language S we feed each possible translation (language I) to the second dictionary, obtaining a set of possible translations in our target language (language T): $\mathcal{D}_1 \circ \mathcal{D}_2 : L_S \mapsto \mathcal{P}(L_T)$. Note that this composition is defined as the composition of $\mathcal{D}_2$ for each word w_I that results from applying $\mathcal{D}_1$ to a specific source words w_S.
- **Addition:** The addition of two dictionaries $\mathcal{D}_1 : L_S \mapsto \mathcal{P}(L_T)$ and $\mathcal{D}_2 : L_S \mapsto \mathcal{P}(L_T)$ results in a dictionary $\mathcal{D}_{1+2} : L_S \mapsto \mathcal{P}(L_T)$ where, for each word w_S from the source language, we compute the union the the possible translations from each dictionary.

Using the same substitution process as described earlier, we obtain the results presented in table 4. With this bigger dictionary, with possibly more errors, we get some more words that maintain orthography between languages, but also more words where substitutions produce valid words. The precision drops from the previous 99% to 96.6% (still above 95%), and the recall from the nearly 80% from the previous evaluation to 68.9%. The F_1 measure keeps above 0.80.

4 Dictionary Alignment

As explained before, our main goal is the alignment of entries from *Dicionário-Aberto* (DA) with the revised edition of the *Diccionario de Sinónimos da Lingua Galega* (DSLG).

One problem with this process is that DA uses an old Portuguese orthography, but some work has already been initiated to modernize its language. Although the Portuguese orthography is changing again (with the late adoption of an Orthography Agreement from 1990 [3]), the process of modernization is being performed to the orthography used before 1990. The main reason is that it is easy to migrate it to the current orthography [1], but the inverse

Table 4 Given 14 492 pairs mapping Portuguese words to a set of possible Galician words, the table presents precision, recall and F_1 measure; accuracy, total of correct words, and delta of correct words from last run. Note that substitutions are cumulative (meaning that when substitution B is performed, substitution A was performed before).

Subst. Id.	Precision	Recall	F_1	Accuracy	Correct	Δ
ID	0.9668	0.5022	0.6611	0.4937	7155	7155
A	0.9664	0.5176	0.6741	0.5084	7368	213
B	0.9663	0.5275	0.6824	0.5179	7506	138
C	0.9668	0.5646	0.7129	0.5538	8026	520
D	0.9661	0.5746	0.7206	0.5633	8163	137
E	0.9658	0.5831	0.7272	0.5713	8279	116
F	0.9658	0.5834	0.7274	0.5716	8283	4
G	0.9656	0.5875	0.7305	0.5754	8339	56
H	0.9648	0.5953	0.7363	0.5827	8444	105
I	0.9648	0.5958	0.7367	0.5831	8451	7
J	0.9649	0.5986	0.7388	0.5858	8490	39
K	0.9654	0.6204	0.7554	0.6069	8795	305
L	0.9656	0.6274	0.7606	0.6136	8893	98
M	0.9656	0.6311	0.7633	0.6172	8944	51
N	0.9662	0.6439	0.7728	0.6297	9126	182
O	0.9661	0.6451	0.7736	0.6308	9142	16
P	0.9662	0.6470	0.7750	0.6327	9169	27
Q	0.9663	0.6542	0.7802	0.6396	9269	100
R	0.9663	0.6556	0.7812	0.6410	9289	20
S	0.9662	0.6631	0.7865	0.6481	9392	103
T	0.9661	0.6657	0.7882	0.6505	9427	35
U	0.9662	0.6719	0.7926	0.6565	9514	87
V	0.9661	0.6730	0.7934	0.6575	9529	15
W	0.9662	0.6735	0.7937	0.6579	9535	6
X	0.9660	0.6746	0.7944	0.6590	9550	15
Y	0.9659	0.6757	0.7951	0.6600	9564	14
Z	0.9659	0.6759	0.7952	0.6601	9566	2
AA	0.9659	0.6762	0.7955	0.6604	9571	5
AB	0.9659	0.6768	0.7959	0.6610	9579	8
AC	0.9659	0.6771	0.7961	0.6613	9584	5
AD	0.9660	0.6781	0.7968	0.6623	9598	14
AE	0.9660	0.6797	0.7979	0.6638	9620	22
AF	0.9660	0.6804	0.7984	0.6644	9629	9
AG	0.9659	0.6814	0.7991	0.6654	9643	14
AH	0.9660	0.6819	0.7994	0.6659	9650	7
AI	0.9661	0.6841	0.8010	0.6681	9682	32
AJ	0.9660	0.6863	0.8025	0.6701	9711	29
AK	0.9660	0.6873	0.8032	0.6711	9726	15
AL	0.9661	0.6881	0.8037	0.6718	9736	10
AM	0.9660	0.6884	0.8039	0.6721	9740	4
AN	0.9660	0.6887	0.8041	0.6724	9744	4

Table 5 Substitution used, the number of words from the Portuguese dictionary with translation (and corresponding percentage), the number of words from the Galician dictionary used as translations (and the corresponding percentage).

	Portuguese Words		Galician Words	
Substitution	Count	Percentage	Count	Percentage
ID	12711	15.3502%	12711	33.7475%
A	13082	15.7982%	13065	34.6874%
B	13447	16.2390%	13421	35.6326%
C	14348	17.3270%	14321	38.0220%
D	14764	17.8294%	14728	39.1026%
E	15174	18.3245%	15138	40.1912%
F	15179	18.3306%	15143	40.2044%
G	15311	18.4900%	15263	40.5230%
H	15856	19.1481%	15807	41.9673%
I	15874	19.1699%	15820	42.0019%
J	15953	19.2653%	15899	42.2116%
K	16365	19.7628%	16306	43.2922%
L	16571	20.0116%	16512	43.8391%
M	16683	20.1468%	16624	44.1365%
N	16716	20.1867%	16657	44.2241%
O	16752	20.2302%	16693	44.3197%
P	16797	20.2845%	16738	44.4391%
Q	16969	20.4922%	16910	44.8958%
R	17003	20.5333%	16944	44.9861%
S	17150	20.7108%	17091	45.3763%
T	17237	20.8159%	17178	45.6073%
U	17359	20.9632%	17300	45.9312%
V	17420	21.0369%	17361	46.0932%
W	17436	21.0562%	17377	46.1357%
X	17469	21.0960%	17410	46.2233%
Y	17505	21.1395%	17445	46.3162%
Z	17505	21.1395%	17445	46.3162%
AA	17511	21.1468%	17451	46.3321%
AB	17521	21.1588%	17461	46.3587%
AC	17524	21.1625%	17464	46.3667%
AD	17564	21.2108%	17504	46.4729%
AE	17586	21.2373%	17526	46.5313%
AF	17596	21.2494%	17536	46.5578%
AG	17647	21.3110%	17564	46.6322%
AH	17669	21.3376%	17584	46.6853%
AI	17712	21.3895%	17627	46.7994%
AJ	17740	21.4233%	17648	46.8552%
AK	17765	21.4535%	17673	46.9215%
AL	17784	21.4764%	17693	46.9746%
AM	17813	21.5115%	17718	47.0410%
AN	17817	21.5163%	17722	47.0516%
DIC	20084	24.2540%	19989	53.0705%

process is not injective. Also, the rules used to translate Portuguese words into Galician words take advantage of the orthography before 1990: they would be harder to write for modern Portuguese as it is more ambiguous.

DA has more than 128 000 entries, but as we are also maintaining words in the old orthography, the number of real different words is lower. Also, as the modernization process is not 100% accurate, and to remove some extra error from this process, we used the *Vocabulário Ortográfico do Português*[4] [2] (VOP) to filter what words to align. The removal of duplicate entries (like *pharmácia* and *farmácia*, where only the latter should be used) results in about 110 000 different entries. The VOP lexicon includes more than 155 000 different words. The intersection of these two lexicons includes 82 807 entries. These are the entries we are trying to align at this moment. Regarding the DSLG lexicon, it has 24 571 entries (41 923 meanings or groups of synonyms), totalling 37 665 unique words (entries or synonyms).

Table 5 presents the results of the alignment process. Although the difference between pure string matching (identity function) and the use of substitutions is not huge if we look into percentages, the truth is that the use of substitutions was able to align about five thousand words, and almost half of the Galician dictionary was used as a translation. The final line of the table (identified as `DIC`) is the result of using the substitutions and, for those words that after being translated do not exist in the target lexicon, using the dictionary used in the second evaluation we performed (section 3.3), resulting in a few more than two thousand words recognized.

The big difference between these two dictionaries, and the fact of their being dictionaries (and therefore including a lot of unfrequent words) explain the low percentage of success. Nevertheless, further research should be done in order to understand how the substitutions set can be made better for bigger result sets.

5 Final Remarks

In this paper we present an approach to translate Portuguese words in a dictionary into Galician words using a set of string substitutions. Although the approach is unable to translate all words (and that was never our goal), it can be used to translate a reasonable amount of Portuguese words with a decent precision value.

Nevertheless, we deliberately ignored a relevant problem: false friends. These are words that have the same or similar writing in Portuguese and Galician, but have different meanings. There are mainly two different situations:

- two words that share a subset of the meanings. For instance, *talho* (PT) and *tallo* (GL) share the majority of their senses, but there are some of them that are specific to Portuguese (for example, the place where meat is sold);
- two words that have complete different meanings. An example would be the word *presunto* (written in the same way in the two languages) that means *ham* in Portuguese (a noun), but means *alleged* in Galician (an adjective);

In the first case the alignment between the two entries should be kept. But the second case is completely wrong, and should probably be removed from the alignments. In order to do that, a list of false friends would be needed, or some kind of heuristic to detect the semantic distance between the dictionary entries. In any case, this research direction should be followed in the near future in order to guarantee a high quality level in the dictionary alignment results.

[4] Portuguese Orthographic Vocabulary

This work should be extended in two different directions: first, researching the results obtained, to understand how a larger percentage of alignments can be achieved (and evaluating the alignment quality); second, analysing how incorporating the Galician dictionary into Dicionário-Aberto can result in a better user experience. For instance, Dicionário-Aberto includes a navigation ontology (relations between concepts are extracted and presented to the user as a navigation feature). It might be possible to use the alignment between the two dictionaries to obtain better concept relations, and therefore a more complete navigation ontology.

Acknowledgments This work was partially supported by Grant TIN2012-38584-C06-04, supported by the Ministry of Economy and Competitiveness of the Spanish Government on "*Adquisición de escenarios de conocimiento a través de la lectura de textos: Desarrollo y aplicación de recursos para el procesamiento lingüístico del gallego (SKATeR-UVIGO)*"; by the Xunta de Galicia through the "*Rede de Lexicografía (Relex)*" (Grant CN 2012/290) and the "*Rede de Tecnoloxías e análise dos datos lingüísticos*" (Grant CN 2012/179); and by The Per-Fide project (grant reference no. PTDC/CLEL-LI/108948/2008, from the Portuguese Foundation for Science and Technology, and co-funded by the European Regional Development Fund).

References

1 José João Almeida, André Santos, and Alberto Simões. Bigorna – a toolkit for orthography migration challenges. In *Seventh International Conference on Language Resources and Evaluation (LREC2010)*, Valletta, Malta, may 2010.

2 Margarita Correia (coord.). Vocabulário Ortográfico do Português, 2010. Lisbon: ILTEC/-Portal da Língua Portuguesa.

3 Diário da República. Acordo ortográfico da língua portuguesa, 1990. Technical Report 193, série I-A, 23 de Agosto, 1991. `http://www.portaldalinguaportuguesa.org/index.php?action=acordo&version=1990`.

4 Mikel Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, pages 1–18, July 2011.

5 Xavier Gómez Guinovart, Xosé María Gómez Clemente, Andrea González Pereira, and Verónica Taboada Lorenzo. Galnet: WordNet 3.0 do galego. *Linguamática*, 3(1):61–67, 2011.

6 Xavier Gómez Guinovart, Alberto Álvarez Lugrís, and Eva Díaz Rodríguez. *Dicionario moderno inglés-galego*. 2.0 Editora, Ames, 2012.

7 Camiño Noia Campos, Xosé María Gómez Clemente, and Pedro Benavente Jareño. *Diccionario de sinónimos da lingua galega*. Galaxia, Vigo, 1997.

8 Alberto Simões and José João Almeida. Processing XML: a rewriting system approach. In Alberto Simões, Daniela da Cruz, and José Carlos Ramalho, editors, *XATA 2010 — 8ª Conferência Nacional em XML, Aplicações e Tecnologias Aplicadas*, pages 27–38, Vila do Conde, Maio 2010.

9 Alberto Simões, José João Almeida, and Rita Farinha. Processing and extracting data from Dicionário Aberto. In Nicoletta Calzolari et al., editor, *Seventh International Conference on Language Resources and Evaluation (LREC2010)*, pages 2600–2605, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

10 Alberto Simões and Rita Farinha. Dicionário Aberto: Um novo recurso para PLN. *Vice-Versa*, 16:159–171, December 2011.

11 Alberto Simões, Álvaro Iriarte Sanromán, and José João Almeida. Dicionário-aberto – a source of resources for the portuguese language processing. *Computational Processing of the Portuguese Language, Lecture Notes for Artificial Intelligence*, 7243:121–127, April 2012.