Source Language Markers in EUROPARL Translations

Hans van Halteren
Radboud University Nijmegen
Department of Linguistics / CLST
P.O. Box 9103, NL-6500HD Nijmegen
The Netherlands
hvh@let.ru.nl

Abstract

This paper shows that it is very often possible to identify the source language of medium-length speeches in the EUROPARL corpus on the basis of frequency counts of word n-grams (87.2%-96.7% accuracy depending on classification method). The paper also examines in detail which positive markers are most powerful and identifies a number of linguistic aspects as well as culture- and domain-related ones.

1 Introduction

The EUROPARL Corpus (Koehn, 2005) is one of the most important resources for translation research. It is used extensively, mostly but certainly not exclusively in statistical machine translation. In much of that research, the relation between a source language SL and target language TL is investigated on the basis of aligned sentences of SL and TL without considering whether SL is indeed the actual source language. In this paper we question the lack of attention for the true SL, because we expect significant differences between texts written originally in TL, texts translated from SL to TL and texts translated from another language into TL, even if all three types have been produced by native speakers of TL.[2] At least with respect to the differences between original and translated texts, our expectations are supported by the literature (for an overview, see e.g. Baroni and Bernardini, 2006).

We embarked on a data-driven investigation, using several text classification techniques to determine if the differences are salient enough to distinguish between different source languages for various texts and target languages. In this paper, we first describe the data and methods we used (Sections 2 and 3). Then we present the classification results (Section 4), after which we give a description of some types of markers we were able to identify (Section 5). Finally, in Section 6, we present our conclusions and plans for future research.

2 Experimental Data

The experimental data was taken from the EUROPARL Corpus, but had to be preprocessed to some degree to be suitable for our experiments. First of all, the annotation of the corpus includes a LANGUAGE attribute, indicating the original language of the text. However, it is often absent in one or more versions of the text. If we examine the speeches of at least 100 words in the whole corpus, we find that 11% of them lack the attribute in all available versions and for 5% the attribute has different values in different versions. We linked the attribute across the various versions and excluded all speeches showing inconsistent values.

We decided to focus on the six most common languages in the corpus: English (EN), German (DE), French (FR), Dutch (NL), Spanish (ES) and Italian (IT). For each of these languages as a source language, we aimed for 1000 speeches (henceforth: *texts*) which were present as original and as translations into each of the other five languages. Furthermore, we decided to focus on the medium range as for length, avoiding very short and very long texts, as these were likely to

[2] In principle, the EU translation service always lets translators translate into their mother tongue. Given that we will here be working only with major EU languages, it is unlikely that the principle was violated for our selected texts.

be different in nature.[3] Our aim of 1000 texts from each source language led us to settle on texts between 380 and 2500 words.

Finally, as we wanted the classification systems to focus on language use rather than contents, we transformed the tokens in the texts, separately for each target language. Only tokens that occurred in at least 10% of the texts remained intact. Other tokens were mapped to a marker <X>.

3 Methods

We classified all texts in our experimental set with several text classification methods (see Sections 3.1 to 3.3), using 10-fold cross-validation, each time with 80% of the texts used to train models for each source language, 10% of the texts used to tune parameters and to base thresholds on and 10% of the texts to test classification accuracy. For all methods, the same train/ tune/test splits were used and some dozens of promising parameter settings were tried, after which the best one for each specific run was selected on the basis of tune set results.

Each method was provided with the same text features, viz. n-gram counts. From each selected text (in its mapped version), we extracted counts for all uni-, bi- and trigrams of tokens occurring in at least 10% of the texts (i.e. those not transformed to `<X>`) and allowing `<X>` markers to intervene between any two tokens in the n-gram. For example, the sequence `"I join with Roy Perry in"` would give rise to the n-gram `<3>_I_<X>_with_<X>_<X>_in`, with the `<3>` indicating that there are three real tokens in the n-gram, and possibly any number of intervening `<X>` markers.

3.1 Marker-Based Classification

Our first classification method attempts to classify texts on the basis of individual markers which are by themselves a strong indication that the text originated in a certain source language. All n-grams which occurred more often in the training data with a specific source language SL than with all other source languages taken together were deemed to be *markers* for SL.

When classifying test data, each marker for an SL observed in the test text leads to an increase of the score for that SL for that text. The exact calculation of the increase depends on some parameter settings. It always involves both the marker's precision (how often its presence indeed coincides with the specific SL) and its recall (how many of the specific SL texts contain the marker), but with a weighting favouring precision over recall by a factor of 10 to 100. Precision and recall can be based on either raw or smoothed counts. The calculation may also involve the frequency of the marker in the test text. In this way, the frequency of a marker is not taken into account for determining whether it is a marker and what value it is given, but may be taken into account when classifying test texts.

3.2 Linguistic Profiling

The second classification system was Linguistic Profiling, which was previously shown to be useful for language verification (van Halteren and Oostdijk, 2004), a task which is similar to the current task. Roughly speaking, it classifies on the basis of noticeable over- and underuse of specific n-grams. As the marker-based classification only used overuse and the necessary degree of overuse is lower for Linguistic Profiling, the latter pays attention to many more features and should be able to attain a better classification rate. However, if we want to determine which features are most powerful, interpreting the workings of Linguistic Profiling will be more difficult than for the marker-based approach.

All n-gram counts were normalized to counts per 1000 words. Furthermore, in order to reduce the number of counts, so that the system could cope with the resulting vectors, we included only n-grams which occurred in at least 10 texts. This led to vectors with about 90,000 counts for each target language.

3.3 Support Vector Methods

Finally, we employed Support Vector Machines, viz. LIBSVM (Chang and Lin, 2001). We offered the same vectors we used for Linguistic Profiling to both standard Support Vector Classification (SVC; RBF kernel, various settings for C and γ) and Support Vector Regression (ν-SVR; RBF kernel, various settings for C, γ and ν).

The Support Vector methods use all available information rather than focusing on over- and underuse. They should therefore attain the best classification results. However, extracting information about salient features from the results will be virtually impossible.

[3] E.g., short texts tend to be interruptions and statements such as the opening of the session, where long texts include presentations of written reports.

	MB	LP	SVC	SVR
SL vs TL: TL	84.1	91.5	92.6	95.2
SL vs TL: SL	75.3	88.4	87.6	91.6
Combination	87.2	94.7	90.1	96.9
SL vs SL: SL	74.2	85.8	85.4	89.7
Combination	81.1	91.8	85.4	94.3

Table 1: Average accuracies for two-way SL vs SL choices for various classification methods. The top part represents cases where one of the SL is the TL, the bottom where neither SL is TL.

3.4 Score Comparability

For all methods except SVC, the ranges of test text scores vary greatly as train sets and parameter settings are changed. As we wanted to combine scores from classifications based on models for various SL, however, we needed the scores to be comparable. In order to make them so, we compared the score for each test text with the scores for all tune texts.[4] All lower scoring positive examples provided a increase of the final (comparable) score and all higher scoring negative examples a decrease. For fine tuning, the relative position of the test text's score between the next higher and next lower scoring tune texts was also taken into consideration.[5]

4 Classification Results

We first used our classification techniques to choose between every possible pair of source languages for each text (Section 4.1). Then we combined the various two-way decisions into a single six-way decision (Section 4.2).

4.1 Two-Way Decisions

For each choice between two possible source languages, there are two classification models (one for each of the SL) that can make the choice individually, but we can also combine the two opinions by choosing the SL whose classification model claims the text with a higher score.[6]

The quality of these decisions, for those cases where the actual SL is present in the pair, is shown in Table 1. We distinguish two types of choices. The top of the table shows the results for the cases in which one of the two SL is the target language, with SL equal to TL indicating that the text is an original TL text rather than a translation. For the cases represented in the table, i.e. cases where one of the two SL is the actual one, the task measured here is in fact translation recognition. As the table shows, the accuracy for this task is higher when modeling original TL text than when modeling translations, meaning that it is easier to spot (violations against) regularities in the target language than it is to spot regularities in the translations from a specific source language. Combining the two models yields even better classification, except for SVC, where combination cannot be used properly because SVC only produces scores of 1 and -1. The bottom of the table shows the results for the cases in which both SL are different from the target language. Here the choices appear to be more difficult on average, but the classification quality is still impressive.

		MB	LP	SVC	SVR
ES		64.7	82.9	81.1	87.4
DE		62.0	81.6	80.8	87.5
EN		60.4	80.6	79.3	86.8
FR		58.7	77.0	77.7	85.7
NL		58.4	76.7	75.2	83.1
IT		52.8	73.6	60.7	81.5
ALL TL		87.2	90.6	91.5	96.7

Table 2: Average accuracies for six-way SL choices for various classification methods. The rows with a TL indication represent the results when the text is only available in that TL. The bottom row shows the combined result, using all six versions of the text.

4.2 Six-Way Decisions

Once we had classification scores for the choice between all fifteen possible pairs of source languages (actually the combination scores), we could use these to choose a single SL from among the set of six possible SL. For the current paper, we did this by simply adding all classification scores in favour of each specific SL and then choosing the SL with the highest total. The addition can either be done over all two-way choices referring to a specific target language, or over all choices for all six possible target languages.

The quality of these decisions is shown in Table 2. For the individual target languages, the

[4] We deliberately chose a non-parametric technique here. However, preliminary additional experiments show that a parametric alternative may actually provide slightly better results.

[5] This technique did not work for SVC, as SVC only scores a text with 1 or -1. However, for SVC the technique was also not needed as the scores were already comparable. We did add a very small random number to each score to resolve ties where necessary.

[6] Remember that we made all classification scores comparable.

	EN	DE	FR	NL	ES	IT
EN	980	1	12	2	1	4
DE	16	961	6	9	3	5
FR	15	3	969	2	7	4
NL	16	10	12	956	4	2
ES	14	5	13	2	960	6
IT	8	2	9	0	8	973

Table 3: Confusion table for SVR classification of each of the 6000 selected speeches on the basis of all six language versions of each speech. Rows are actual languages; columns are languages assigned by the classifier.

accuracies are not all that high, but this was to be expected. Interestingly, the relative difficulty of the choice for the various TL is largely independent of the classification method used, with texts in Spanish being most easily classifiable and texts in Italian least easily. The relative performance of the classification methods is again as expected, except that SVC is slightly behind Linguistic Profiling for the individual languages.

For the best performing classification method, Support Vector Regression, we also show the confusion table for SL assignment (Table 3). In general, there are more confusions within the Germanic and Romance language families than between them. However, English seems to take an intermediate position between the two families. On the other hand, the English model appears to be somewhat greedier than the others, which confuses the analysis. Also, given that the LANGUAGE attribute does not seem to be assigned perfectly, it would be advisable to examine the misassigned texts, so as to check if they are indeed misassigned or merely mislabeled.

5 A Look at Source Language Markers

Now we have shown that word n-grams provide a solid basis for source language identification, we proceed to an examination of the n-grams that are the most useful in this identification. Unfortunately, as stated above, Linguistic Profiling and the Support Vector methods are not very amenable to extracting information about the most salient features. Therefore, for now, we will have to fall back on the marker-based classification where it is trivial to identify the source language markers which characterize their source languages most strongly. Admittedly, the marker-based classification had the lowest performance in the classification task, but still it is good enough for an examination to make sense.

	Occurring only once; With SL	Occurring more often; Only with SL	Occurring also with non-SL
DE	137256	6691	9459
FR	122386	4916	7102
NL	120071	5594	7653
ES	119899	5740	8495
IT	129251	5900	8538

Table 4: Numbers of source language markers for various source languages SL in translation to English.

To optimize understandability for all readers, we focus on English as the target language in this examination. We distinguish between two kinds of marker strength. Obviously, there is the strength in the SL classification described above. For a marker to be strong in this sense, it has to occur more often with the source language in question than with all other source languages taken together (Section 5.1). If, however, we envision an application where we know that a specific SL is being translated to a specific TL and we want to give feedback to the translator that the translation contains strong influences from the SL, then strength can also be taken to be the degree to which the marker occurs more in texts translated from SL than in original TL texts (Section 5.2). For both kinds of marker strength, we can identify specific types of markers, related either to linguistic or culture- and domain-related aspects of texts and their translations (Section 5.3)

5.1 Statistics for SL vs All Others

We took all n-grams from the English versions of the texts in our experimental data and counted the number of texts they were contained in for each of the six source languages. We then calculated their strength for each source language SL by dividing the observed number of SL texts, plus one, by the number of non-SL texts, again plus one. The classification described in Section 3.1 in principle used all markers with strength greater than one.

If the full set of texts were used as training material, we would find the numbers of markers as shown in Table 4. The majority of n-grams occurs only once (Column 2) and necessarily with a single specific SL so that they can be taken to be markers. Note, however, that these markers did not play a role in the classification in the ten-fold cross-validation, since they were always either only in the training set, only in the tune set or only in the test set. The third column represents n-grams which also occur only with

SL	n-gram	texts
DE	<3>_is_,_though	10
	<2>_Commission_here <3>_,_again_and <3>_are_right_<X>_you <3>_If_,_though <3>_me_say_this <3>_the_Commission_here <3>_the_<X>_Council_Presidency	7
	<3>_._What_that <2>_What_that <3>_reason_we_must <3>_must_at_last <2>_more_able <3>_go_without_saying <3>_not_,_though <2>_need_<X>_- <2>_taken_here <3>_believe_that_here <3>_is_needed_here <3>_gentlemen_,_<X>_<X>_is	6
FR	<3>_common_to_the <3>_quality_of_her	6
	<3>_with_no_regard <3>_no_regard_for <2>_no_regard <3>_conditions_of_<X>_of <2>_into_<X>_this <2>_on_<X>_services <2>_particular_about <2>_various_policies <3>_all_those_, <3>_of_cooperation_is <3>_the_United_<X>_<X>_Commissioner <3>_,_<X>_society_,	5
NL	<3>_like_to_<X>_off	23
	<3>_to_<X>_off_with	16
	<3>_On_a_final	14
	<2>_Commissioner_whether <3>_the_Commissioner_whether	10
	<3>_and_such_like <2>_such_like	8
	<3>_past_<X>_of_years <3>_,_it_<X>_as <3>_are_in_order <3>_too_<X>_for_words <3>_to_<X>_off_by	7
ES	<3>_going_to_support	11
	<3>_amendments_presented_by <3>_end_here_, <3>_Community_system_for <3>_the_people_responsible	8
	<3>_citizens_._And <3>_going_to_debate	7
	<3>_the_Community_<X>_sector <2>_Community_<X>_sector <2>_than_<X>_<X>_<X>_million <3>_million_<X>_year_. <3>_President_,_without <3>_Let_us_see <3>_adopt_measures_to <3>_move_<X>_with_the <3>_And_the_Commission <3>_people_responsible_for	6
IT	<3>_least_in_that	9
	<2>_task_before <3>_or_,_<X>_still	7
	<3>_change_the_current <3>_is_the_Europe <3>_feel_that_Parliament <3>_task_before_us <3>_of_my_Commission <2>_with_<X>_<X>_' <3>_the_<X>_available_. <2>_:_<X>_which <3>_security_and_peace	6

Table 5: Strongest n-grams occurring only with a specific SL

one specific SL. These markers are in principle the most useful because of their precision. However, their number of occurrences turns out to be generally very low, with a maximum of 23 and a mean of 2.1, so that their recall is low to very low. Finally, there are the n-grams which do occur with various source languages, but most often with one specific SL. They are represented by the last column.

Probably more insightful than the general statistics are the actual markers themselves, here represented by the strongest ones for each source language, still with English as the target language. The lists we find are always headed by n-grams which occur only with a specific SL (Table 5). Then follow the n-grams which can also be found with other languages. We show the strongest ones of this type separately as Table 6.

SL	n-gram	SL texts	other source language texts
DE	<3>_means_is_that	11	1 IT
	<2>_framework_conditions	22	2 EN, 1 FR
	<3>_in_future_be	14	1 EN, 1 IT
	<3>_,_that_being	13	2 NL
	<2>_action_here	8	1 FR
	<3>_So_let_me	8	1 FR
	<3>_to_at_last	8	1 FR
FR	<3>_why_I_shall	8	1 NL
	<2>_certain_number	25	1 DE, 2 IT, 2 NL
	<3>_certain_number_of	25	1 DE, 2 IT, 2 NL
	<3>_a_certain_number	24	1 DE, 2 IT, 2 NL
	<3>_thank_our_rapporteur	7	1 NL
	<3>_We_now_know	6	1 NL
	<3>_provide_itself_with	6	1 ES
NL	<3>_._<X>_to_say	61	4 EN, 2 DE, 1 FR
	<3>_that_is_concerned	10	1 DE
	<3>_we_as_Parliament	9	1 IT
	<3>_group_,_it	9	1 ES
	<3>_is_every_reason	9	1 ES
	<3>_deal_of_support	8	1 DE
	<3>_think_that_that	12	1 ES, 1 FR
ES	<3>_._<X>_this_context	9	1 IT
	<3>_I_<X>_this_to	9	1 EN
	<2>_people_responsible	8	1 NL
	<3>_going_to_deal	8	1 NL
	<3>_President_,_<X>_allow	8	1 IT
	<3>_legislation_in_force	8	1 FR
	<3>_report_,_since	8	1 FR
IT	<2>_the_now	10	1 FR
	<3>_other_,_there	10	1 DE
	<3>_the_individual_States	10	1 DE
	<3>_,_<X>_<X>_,_ladies	10	1 DE
	<2>_my_Commission	8	1 FR
	<3>_due_regard_for	12	1 EN, 1 NL
	<2>_quite_aware	16	3 FR

Table 6: Strongest n-grams occurring also with other languages

As already stated above, most of the stronger markers occur in only few SL texts. The big exception seems to be Dutch, which has several markers (notably <3>_._<X>_to_say with competing languages and <3>_like_to_<X>_off exclusively) which show up quite often. Only <2>_framework_conditions for German and the certain_number cluster for French occur over 20 times too, but they are not exclusive. We also see various clusters, where longer strings are represented by several n-grams, such as "a certain number of" and "with no regard for" for

token	EN	DE	FR	NL	ES	IT
<1>_ladies	11	574	362	197	273	378
<1>_gentlemen	14	584	383	205	304	388
<1>_And	69	154	160	164	307	128
<1>_above	58	151	119	115	145	198
<1>_guarantee	53	97	110	87	164	151
<1>_favour	75	157	197	173	164	122
<1>_everyone	55	144	141	126	57	93
<1>_namely	54	158	94	184	54	60
<1>_opinion	140	164	244	272	250	311
<1>_mention	65	109	120	86	127	121
<1>_although	95	122	110	116	233	219
<1>_therefore	296	396	488	477	580	525
<1>_regard	174	212	292	254	408	286
<1>_shall	108	178	240	144	166	169
<1>_though	80	215	131	116	82	117
<1>_everything	69	152	126	98	97	98
<1>_various	107	168	188	154	186	175
<1>_hand	103	160	183	178	145	171
<1>_freedom	72	118	119	88	105	157
<1>_Office	73	158	110	77	142	102

Table 7: Strongest individual tokens marking translations into English

French and "the people responsible for" for Spanish.

5.2 Markers for Translation vs Original

If we want to call a translator's attention to a translation which deviates from the general language use in the target language, the marker identifying the deviation need not be exclusive to a single source language. Such markers are much more common than source language specific markers. They also include large numbers of single tokens (i.e. unigrams; Table 7), something we did not find in Section 5.1.

Most remarkable are the top two, ladies and gentlemen. Apparently, speakers from all over Europe address the whole house when opening their speech, but those speaking English only address the President (i.e. the chairperson). The rest are mostly words providing discourse functions, either by themselves, such as therefore and though, or in larger combinations, such as hand in "on the one hand" and opinion in "in my opinion". There are only a few words with actual content, such as guarantee, freedom and Office, although the latter may also well be part of a parliament term.

As for the longer n-grams, we will not present the full list of strongest markers, but instead a filtered selection (Table 8). The reason for this is that there is quite a lot of repetition in the full list. The strongest eleven, and fifteen more of the strongest fifty, are (parts of) combinations of vocatives, such as "Commissioner, ladies and gentlemen", which we already addressed above. As vocative use in the European Parliament could well be a separate study in itself, and is probably not something we need to bother translators with, we leave out all vocatives.

token	EN	DE	ES	FR	IT	NL
<3>_of_the_Group	0	27	41	19	16	34
<2>_end_by	0	2	42	18	25	7
<3>_,_by_means	0	13	40	21	11	8
<3>_s_<X>_(_<X>_<X>_)	0	35	13	8	11	21
<3>_)_and_European	0	29	14	8	11	21
<3>_,_for_we	0	25	0	7	26	20
<3>_at_last_,	0	9	3	5	53	5
<3>_countries_,_which	0	13	16	24	10	9
<2>_we_therefore	0	10	19	8	22	10
<3>_the_Europe_of	0	5	12	18	30	3
<3>_order_to_guarantee	0	9	26	14	6	11
<3>_and_above_all	1	35	29	26	28	13
<3>_out_that_,	0	11	22	12	12	8
<2>_therefore_believe	1	13	67	21	19	8
<3>_think_that_this	1	14	22	29	15	46
<3>_think_that_,	0	9	15	14	10	11
<3>_like_to_<X>_this	0	5	15	4	7	27
<3>_of_third_countries	0	4	17	15	12	9
<2>_here_too	0	27	1	4	8	15
<3>_of_all_like	0	4	12	14	2	22
<3>_,_though_,	3	128	4	17	14	53
<3>_too_,_we	0	21	0	6	11	15
<3>_I_shall_not	0	3	10	21	7	12
<3>_State_or_Government	0	15	5	10	16	6

Table 8: Selection from the strongest n-grams marking translations into English

5.3 Types of Markers

The markers shown in the previous sections are of a rather varied nature. Some of them have linguistic explanations, but there are also quite a few which are more culture- and domain-related.

The best example in the latter category is the already mentioned use of vocatives. Although there are clear links to at least one source language, English, this is not something that is caused by translation. Another example of seemingly typical parliamentary behaviour are the phrases "like to finish off" and "On a final note", responsible for the three strongest Dutch markers in Table 5. The Dutch (or Flemish) parliamentarians announce in some way that they are nearing the end of their speech. However, if we examine the original Dutch text, we observe a much more varied phrasing. We find the literal counterpart ("ik wil afsluiten"), but also "ter afsluiting" ("to close") and "dan nog iets" ("then another thing"). Apparently, one or more of the Dutch to English translators have developed their own favourite phrases to cover this general situation.

Another domain-specific type of marker can be found in content words (here mostly compounds) referring to parliamentary matters. Here we turn to German for some examples. In Table

5 we find <3>_the_<X>_Council_Presidency. One might think that only Germans are interested in who runs the council at any given time, but in fact this is an idiosyncratic alternative translation as "Council Presidency" is generally just called Presidency. Another example is <2>_framework_conditions in Table 6, with 22 German SL occurrences, 2 English and 1 French. When examining the source text, we find Rahmenbedingungen, elsewhere translated (probably better) as "basic conditions". From these examples it would seem that some work may still be needed on harmonizing terminology.[7] A lack of (knowledge of) central terms leads in one case to a translation with some unfortunate connotations and in another case to an acceptable but deviant translation. In both cases the use of a single term would most certainly also improve information retrieval on the parliament proceedings.

Related to the domain, but much more culture specific is the variation in the way the speakers organize their argumentation. The example of speakers of Dutch announcing the last part of their speech has already been mentioned. Another thing speakers of Dutch seem to do is that they exaggerate their viewpoint, both positively and negatively. The positive exaggeration is visible in the word natuurlijk (naturally, obviously, …), which is found in almost a third (326) of the originally Dutch speeches. One of the translations chosen for this word is "needless to say", thus giving rise to the extremely strong marker for Dutch in Table 6. The negative exaggeration is present in cases where a situation is called insane: waanzinnig, "te gek voor woorden" ("too crazy for words") or "te gek om los te lopen" ("too crazy to walk around freely"). A possible translation template here is "too crazy/absurd/ridiculous for words", explaining another marker for Dutch in Table 5. A further obviously discourse-related marker is therefore. As Table 7 shows, it is found from 30% (DE) to 100% (ES) more in translations than in original English text. This may mean that speakers of the other languages place more causal relations in their arguments, but it can also be that therefore just happens to be a favourite translation option among the translators. Again, this could be a research topic by itself and a thorough investigation is beyond the scope of this paper.

Obviously, each source language has some words or phrases in its vocabulary which make themselves felt in the translations. Apart from the ones already mentioned, the strongest example here is the French marker "a certain number of". The original turns out to be "un certain nombre", elsewhere more English-like translated as "some of".

There must also be instances of influences from languages' syntax, but these are much harder to find. It is possible that the overuse of shall, especially for French, is linked to verbs which are morphologically marked for future tense. Also the overuse of And at the beginning of sentences may be linked to splitting source language long sentences into two English sentences or merely to more extensive use of a coordinating connective in the other languages. For both these words, there are also far too many occurrences to examine at this time.

6 Comparison to Other Work

The most similar investigation we are aware of is that by Baroni and Bernardini (2006). They worked on a corpus of Italian geopolitical journal articles and used SVMs to distinguish translated and original Italian text on the basis of mostly n-gram features representing both types of text. They did not attempt to identify the source language. Their task corresponds to the translation recognition task presented in the top half of Table 1 and their method is comparable to the combination of the two models for original texts and translations. They report an accuracy of 86.7%. If we examine the texts with TL equal to Italian, we find combination scores of 85.4% (MB), 86.0% (SVC), 93.2% (LP) and 96.3% (SVR). Although their work is different in the choice of domain (geopolitical journal articles) and they do not distinguish translated texts as to source language, the results for SVM classifiers are comparable.

They also mention that in earlier research (Baroni and Bernardini, 2003), they found that *"bigrams most characteristic of translated text are sequences of function words"*, for both the Italian corpus already mentioned and a corpus of EU reports written in and translated into English. The 2003 paper itself, however, reports that *"a more thorough investigation of the EU data … failed to reveal systematic differences between translated and original documents"*. We must

[7] Since EUROPARL only contains speeches some years in the past, the situation may well have been remedied in the meantime.

therefore conclude that their observation must refer to the Italian texts. Still, our tables do show quite a few function word bigrams, but of course we have contributed to any such predominance by blanking out most content words.

Borin and Prütz (2001) examined translations from Swedish into English, using news articles. They examine over- and underuse of POS n-grams. They manage to explain some of their observations, but not the overuse in the translations of adverbs, infinitives, pronouns and sentence-initial prepositions. We have not examined POS classes, but only specific words. However, we do see various adverbs in prominent positions, especially in Table 7, which indeed shows overuse. The sentence-initial prepositions might also be partly explained by (often lexicalized) adverbial prepositional phrases.

7 Conclusions

We have shown that classification on the basis of word n-grams markers is able to identify the source language of medium-length European Parliament speeches. Depending on the classification method used, the actual source language can be identified for 87.2% to 96.7% of the texts when all six target language versions of the text can be accessed. If only a single version is available, classification is considerably worse and only Support Vector Regression consistently shows relatively high scores, with accuracies of 81.5% for the Italian rendering to 87.4% for the Spanish one.

We also examined the strongest markers. We found that they are rather varied in nature and represent a wide range of information sources. Vocabulary, discourse structure and probably syntax of the source language all contribute. Contrasts between source and target languages can be seen to have an influence too, both purely linguistically and through the behaviour of the translators. But also the behaviour patterns of the parliamentarians of the various countries have a clear influence. Some of these influences are harmless or even attractive. Others should be followed up on, e.g. it would be good to attempt a harmonization of terminology throughout the various translation services, so that information retrieval on the parliamentary proceedings can be improved.

As for further research, it is vital to first investigate how exactly the European Parliament proceedings have been translated in the past, are being translated in the present and will be translated in the future. It may well be that some of the effects we are finding are limited to individual translators, or caused by their use of (machine) translation tools. Once it is clear that we are measuring what we think we are measuring, namely general trends in speaker and translator behaviour, we need to automate the retrieval of target language and source language phrases (maybe using statistical machine translation methodology), and possibly also to address semantically related clusters. Only then can we really investigate if observations like *"Dutch speakers exaggerate more often"* are valid or are just false impressions from looking at the data through too small a window.

Once all this is in place, we will have the means for a whole range of activities, e.g. to study parliamentary behaviour, to study the translation process, to determine if source language should be taken more into account in EUROPARL translation models and potentially even to give useful advise to the EU translation services.

References

Marco Baroni and Silvia Bernardini. 2003. A Preliminary Analysis of Collocational Differences in Monolingual Comparable Corpora. *Proc. Corpus Linguistics 2003, Lancaster, UK.*

Marco Baroni and Silvia Bernardini.2006. A New Approach to the Study of Translationese: Machine-Learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21(3): 259-274.

Lars Borin and Klas Prütz. 2001. Through a glass darkly: Part of speech distribution in original and translated text. *Computational linguistics in the Netherlands 2000.* Edited by Walter Daelemans, Khalil Sima'an, Jorn Veenstra, Jakub Zavrel. Amsterdam: Rodopi. 2001. 30-44

Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Hans van Halteren and Nelleke Oostdijk. 2004. Linguistic Profiling of Texts for the Purpose of Language Verification. *COLING 2004, Geneva*: 966-972.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit X, Phuket, Thailand*: 79-86.