Automated classification of unexpected uses of this and that in a learner corpus of English

Thomas Gaillat, Pascale Sébillot, Nicolas Ballier

▶ **To cite this version:**

Thomas Gaillat, Pascale Sébillot, Nicolas Ballier. Automated classification of unexpected uses of this and that in a learner corpus of English. Lieven Vandelanotte; Kristin Davidse; Caroline Gentens. Recent Advances in Corpus Linguistics: Developing and Exploiting Corpora, 78, Brill, pp.309-324, 2014, Rodopi Language and Linguistics Special E-Book <10.1163/9789401211130_015>. <hal-01058760v2>

HAL Id: hal-01058760

https://hal.archives-ouvertes.fr/hal-01058760v2

Submitted on 19 Feb 2018

Automated classification of unexpected uses of *this* and *that* in a learner corpus of English

Thomas Gaillat
University of Paris-Diderot and University of Rennes 1
Pascale Sébillot
INSA de Rennes
Nicolas Ballier
University of Paris-Diderot

Abstract
This paper deals with the way learners make use of the demonstratives *this* and *that*. NLP tools are applied to classify occurrences of native and non-native uses of the two forms. The objective of the two experiments is to automatically identify expected and unexpected uses. The textual environment of all the occurrences is explored at text and PoS level to uncover features which play a role in the selection of a particular form. Results of the first experiment show that the PoS features *predeterminer* and *determiner*, which are found in the close context of occurrences, help identify unexpected learner uses among many occurrences also including native uses. The second experiment shows evidence that the PoS features *plural noun* and *coordinating conjunction* influence the unexpected uses of the demonstratives by learners. This study shows that NLP tools can be used to explore texts and uncover underlying grammatical categories that play a role in the selection of specific words.

1 Introduction

In this paper, we present the results of two experiments designed to test automatic unexpected use classification of *this* and *that* in learner English. The objective is to analyse the textual environment of the demonstratives so as to uncover PoS or token-related features which play a role in the selection of a particular *expected* or *unexpected* form of *this* or *that*. This work is part of a wider project in which we intend to annotate learner English, as it is clear that learner-error analysis requires such a type of information (Granger 2008). We want to achieve this automatically by following de Haan (2000) on the ICLE[1] corpus. Our approach follows the path set by Swiss linguist Frei (1929: 25) who insisted on the necessity to see errors as traces of language needs that have to be fulfilled by the learner. We endorse this view in which language facts are to be explained rather than just be compared with a norm. By using corpora and NLP tools as resources to validate or not linguistic explanations of particular language issues, we follow Frei's functional approach of linguistics.

Our work builds on previous work in several domains. Error tagging of learner corpora (Dagneaux et al. 1998) (de Mönnink 2000) has shown that learner English requires specific processing, be it manual or computer-assisted. However, the increasing amount of data makes the manual annotation task a burden. In parallel, NLP tools have been playing an increasing role in annotation issues. Native corpora have been the target for automatic PoS and syntactic annotation for quite some time, with results showing accuracy of around 95% for the former type (Schmid 1994). It was not long until learner corpus was also automatically annotated at PoS level (de Haan 2000) (van Rooy and Schafer 2003). Recently, machine learning technologies have been applied to automatically detect various types of errors such as article selection (Han et al. 2006) (Pradhan et al. 2010). Our approach follows the same line of research by using automatic PoS-tagging annotation for automatic classification on a learner corpus. Instead of article selection, the focus is placed on the selection of demonstratives.

The objective of the paper is to describe two experiments carried out to discover the linguistic characteristics that influence the use of *this* and *that* in expected and unexpected contexts. The principle is to pass on a representation of contexts to a classifier in order to simulate the selection process of a *this* or a *that*. Our hypothesis is that the selection of *this* or *that* depends on its close context of utterance composed of PoS and words called tokens in this article (Cornish 1999: 68). We build a representation of contexts by converting PoS and text into a set of features that a classifier categorizes in classes according to metrics predetermined in a training phase. The first experiment is an attempt to measure the impact of

[1] The International Corpus of Learner English whose director is Sylviane Granger. http://www.uclouvain.be/en-cecl-icle.html.

certain distributional features of the demonstratives on their classification as *expected* or *unexpected* forms.[2] It discriminates *expected* uses of the two forms without distinction against *unexpected* uses of the same two forms. By selecting specific PoS-tags from surrounding contexts of *expected* and *unexpected* occurrences, a classifying process is implemented to see whether or not these selected features play a role in the distinction between *expected* and *unexpected* uses. The second experiment's novelty lies in the nature of the dataset, as only *unexpected* uses of the demonstratives, in their close context, are considered. By using an automatic classifier with this dataset, the goal is to see what specific linguistic features play a role in the classification process of just *unexpected* uses of *this* or *that*. In other terms, the point is to see if the set of features in *unexpected* contexts helps to predict a particular *unexpected* form. The second section of the paper covers the method used to prepare the datasets. The third section deals with the way features are selected and extracted from texts. Finally, the last section discusses the results.

2 The dataset

In this part we describe the two components of the dataset and we explain how they are used in relation to the two experiments.

2.1 Native corpus subset

We create a dataset made up of occurrences of *this* and *that* which come from two corpora. A first subset consists of forty occurrences from the Penn Treebank-tagged Wall Street Journal corpus (Charniak et al. 1987) which are extracted thanks to Tregex (Levy and Andrew 2006). The objective of the extraction process is to obtain twenty singular occurrences of each form, together with their close context composed of token PoS-tag pairs. The forty occurrences are selected randomly. The choice of the WSJ reflects the quality of its PoS tag accuracy as the error rate is estimated at 3% (Marcus et al, 1993). A previous study[3] involved the creation of new PoS-tags for *this* and *that*, so as to provide

[2] The term *unexpected* was favoured over the term error after tests on natives. Non-native occurrences were shown to natives. The tests consisted in presenting actual non-native utterances to natives with gaps replacing *this* and *that*. Natives were first asked to fill the gaps. When their choice contradicted the non-native choice, they were asked to judge the non-native choice. The tests showed that natives would classify choices in three categories: acceptable, unacceptable and acceptable as a second choice. The term *unexpected* covers both the unacceptable and second-choice categories.

[3] This study focused on the detection of features that lead to the selection of demonstratives in their pro-form use in native English. It was done in collaboration with Detmar Meurers (University of Tubingen) and Nicolas Ballier (University of Paris-Diderot).

more accuracy in the distinction of the forms. The tagset includes a clear distinction between the determiner and pro-form functions of the demonstratives. The choice of this corpus also reflects the need to have a good reference point with the learner corpus. Michael Barlow (2005: 345) points out the problem of multiple genres in corpora: “the combination of genres in the general corpus does not provide a good reference point for the learner corpus, which invariably consists of a single genre”. So, the WSJ provides a single genre with which other corpora may be compared. Even if it is a written corpus, single genre and reliable PoS annotation appear as strong factors for the choice of this corpus in the experiments.

2.2 Learner corpus subset

The second subset of the data is an extract from Charliphonia, the University of Paris-Diderot's subset of the Longdale[4] corpus initiated by the Centre for English Corpus Linguistics at the University of Louvain. This corpus is composed of audio recordings of English learners. Learners were interviewed by a native speaker, and asked a few general questions on their recent background. Learners answered these questions exhaustively without many interruptions from the natives. For our experiments forty occurrences of *unexpected* uses of *this* and *that* were identified manually and extracted from the transcripts. Each occurrence was selected with its surrounding context. When the context included an occurrence of a demonstrative which was expected, the context was shortened so as to neutralise any expected use of the form. Without neutralisation, expected uses of the form would also be processed and, thus, introduce a bias to the homogeneity of the dataset. This sample includes 20 occurrences of *this* and 20 occurrences of *that*, which correspond to the two grammatical functions described in the previous paragraph. As the sample is small and in order to avoid variability due to number, only singular occurrences were selected. Consequently, only 40 occurrences of singular forms were selected randomly in the WSJ. The selection of *unexpected* uses was performed manually, and cross-validation was carried out with a native English speaker. A form was characterised as unexpected when the native speaker considered the choice of the demonstrative as not being the obvious one. A previous study[5] shows that alternatives would have been substitutions with the other demonstrative or with the pronoun *it* or the determiner *the*. In other terms, unexpectedness is due to two trends: either the learners swap the two words or they swap the form with an erroneous use of *the* or *it*.

[4] http://www.uclouvain.be/en-cecl-longdale.html.

[5] In a paper (Gaillat 2013), presented at the conference “Learner Corpus Research 2011” in Louvain, we showed that, in learner use, interferences exist within the determination system as the demonstratives compete with the article *the*. Interferences within the anaphoric system also exist as demonstratives compete with the pronoun *it.*

The following examples show the diversity of uses that may be classified as unexpected. In the first one, *this* and *that* are used as pro-forms and refer to the entity pizza. Native informers consulted on this issue would favour the pronoun *it* in both cases.

(1) DID0115-S001 "<A> would you consider pizza an Italian food </A> <B> (em) yes but it's not it's not really f= it's typic but it's not (em) we can eat *that* everyday everywhere now and . but (em) my grandma does *this* by herself"

In the second example, the demonstrative is used in the position of determiner. Native speakers would clearly favour the use of *the* or *this*. The choice of *that* creates a local context of rejection as if the country was of no interest to the speaker. The broader context proves the opposite as the speaker insists on her motivation to live in this country.

(2) DID0145-S002 "[I suppose I'll go to Peru because this is a country I always (er) have intrigued me (er) my mummy my mom gave me a necklace with the God of Sun . Inti and since I had this like four five years ago I have always wanted to go there and to climb the Machu Picchu . so I will I want to go there so .] I'm gonna there for sure (em) . I will .. I will go there for a year I think (er) to work there to help people there . and to discover *that* country . because there are a lot a lot of things to: to find ."

In the last example, *this* is used as a determiner. Unexpected use can even be classified as an error as the agreement between the form and its noun is not respected.

(3) DID0074-S001 "<B> (er) sports (em) not no sports but (em) music (em) because they they (em) in in *this* countries (em) (er)"

2.3 Corpus subsets for the experiments

For the first experiment, we use both subsets described above as we want to see what features lead to the distinction between learner-corpus demonstratives, characterised by *unexpected* uses, and native-corpus demonstratives. This would allow the identification of PoS and token elements that differentiate *expected* from *unexpected* uses. We do not distinguish between *this* or *that* at this point, but we introduce a balance number of the forms in the samples extracted so that classification is not influenced by weight differences. So, the dataset for experiment 1 is composed of two subsets. The even number of occurrences of *this* and *that* forms in each subset gives a 50/50 baseline with which classification can be compared. The small size of the sample is due to the slow process of identifying *unexpected* uses manually. The classifying method explained below takes this into consideration so as to maximise training and test data.

The second experiment is an insight into *unexpected* learner English only. This is why, in a similar approach to (Pradhan et al. 2010), the dataset is only composed of the Charliphonia subset described in 2.2. In other terms, only *unexpected* uses are taken into consideration and the classification process is carried out so as to have a closer insight into the actual selection of a particular *unexpected* form. The idea is to identify what features lead to the selection of a particular unexpected form.

3 Features

In this section, we describe the way an abstract representation of the context is carried out so that the classifier can process the data. We show how the selection of features depends on linguistic criteria. The conversion of these criteria into features for the classifier and the classifying process itself end the section.

3.1 Selection of linguistic characteristics

For the purpose of our experiments, we need to isolate the relevant characteristics in order to convert them into features for the classifier. All the literature on the subject identifies a number of notions that constitute characteristics in the uses of the forms. Biber et al. (1999: 347) distinguish the use of demonstratives according to the notion of distance: "In addition to marking something as known, the demonstrative forms specify whether the referent is near or distant in relation to the addressee". Stirling (2002: 1504) endorses the same vision and adds a distinction between the dependent and independent uses of the demonstratives together with their deictic and anaphoric uses. Halliday and Hasan (1976: 56-68) encompass the same notions and integrate them within the system of endophoric and exophoric reference. Fraser and Joly (1979: 114) follow Hasan and Halliday in their vision of the system of reference, and go further in their analysis of the two forms. Anaphoric and deictic uses of the forms are distinguished within the reference system, and they introduce the notion of speaker's sphere. The notions of conclusion, rejection, distantiation and rupture are usually accompanied with *that*, while the notions of speaker's sphere, identification to the situation, proximity, personalisation, temporal location are usually found with *this.*

The challenge is thus to determine PoS and words that could correspond to these characteristics depending on context. For both our experiments, we choose native English as a reference to establish a list of the characteristics to be tested. Even in the case of the learner-corpus subset, we consider native-English characteristics as relevant since learners target the native language model when expressing. Table 1 shows the linguistic characteristics that have been selected according to their linguistic values and the PoS tags to which they correspond.

Table 1. Candidate features for expected uses and their linguistic justifications

Description and characteristics	**Features (PoS tags[6])**	**Context**
Verb in the preterite form in order to mark temporal distantiation within the context	VBD	Left
Punctuation in order to mark pauses and the speaker's attitude possibly signalling insistence or change of topic	PUNC*	Left + right
Personal pronouns in order to mark the speaker's existence	PRP	Left
Nouns in order to mark the possible co-reference of the demonstrative and the noun	NN	Left + right
Verb in order to mark predicate introduction, and thus of a reference to an entity	VB	Left + right
Wh- adverbs (where, when) and adverbs (not, never) that may mark the existence or rejection of detailed information on an entity	(W)RB	Left + right
Cardinal numbers such as pro-form *one* may mark the need to express contrast, i.e., “*this* one”, “one of *these*”	CD	Left + right
Coordinating conjunction in order to mark possible changes of focus, reference or topic	CC	Left
Complementizer and relative pronoun that may mark the existence of detailed information on an entity	TCOM* / TREL*	Left
Determiner and pre-determiner in order to mark already existing determination of an entity	DT / PDT	Left
Modal in order to mark the speaker's possible attitude in relation to the situation of communication	MD	Right
Preposition in order to mark possible introduction to entity reference	IN	Left + right
Pro-form or determiner in order to mark the category of *this* or *that* for a particular instance	TPRON* / DT	Right

As far as words are concerned, they have been chosen according to several semantic groups that correspond to the notions identified above. Notions such as rejection (i.e., *no*, *never*), foreground/background information and interest (i.e.,

[6] Penn Treebank scheme except when there is an asterisk mark.

want, *hope*, *say*, *tell, first, second*), topic continuity/discontinuity (i.e., *after*, *however*, *then*) support the introspective choice of specific words. In addition, given the fact that the demonstratives are part of the domain of deixis, it was decided to include the words that provide referential information made by the speaker (i.e., *here, there, this, that, now*). The list of words also includes tokens expected to be found next to *this* or *that*, i.e., *'s, all, like, of.*

As the second experiment only deals with unexpected uses in learner English, we think it is important to select specific linguistic characteristics for this experiment and add them to the afore-mentioned characteristics. Based on professional experience with learners, several characteristics are proposed. Experience in correcting both oral and written productions of students helped with the identification of grammatical issues that are found repeatedly amongst students. Firstly, the PoS tag giving information on the existence of plural nouns (NNS) in the right context is isolated, as it would help to determine agreement errors. Secondly, several words that are usually accompanied by learner difficulties are also isolated: *for, since, despite, (in) order, (in) spite* can all be part of direct translations from French, and as such, may appear in *unexpected* uses. The verb *is* is also isolated, as combinations with the demonstratives are not that clear for learners. In all these cases, learners make typical mistakes and the introduction of the words is an attempt to capture the environment in which errors with *this* or *that* occur. For example, some learners tend to say “In order this happen”. By selecting the word *order* as a feature for the classifier, the idea is to see whether it helps with the improvement of the error classification process.

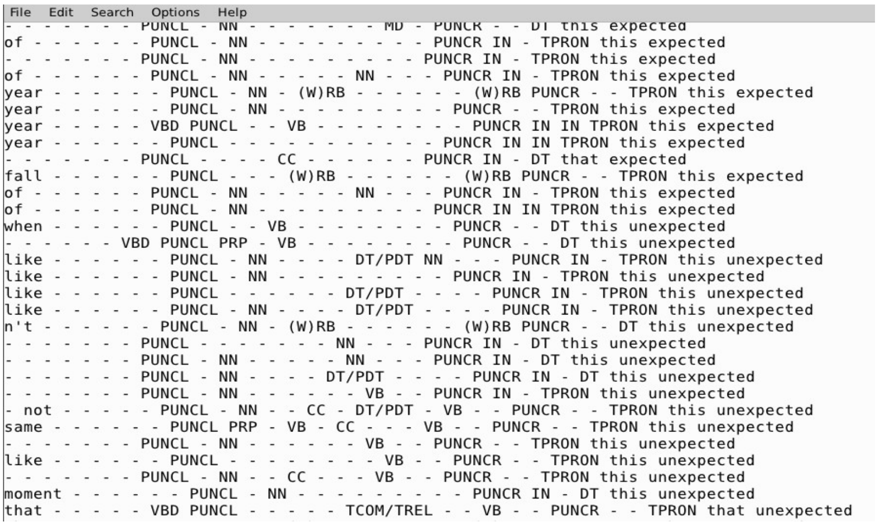
It is important to specify that no one feature can be seen as leading necessarily to the choice of a particular form. Instead, the experiment aims to test whether all the features, as a whole, have an influence or not on the choice of *this* or *that*. At this point in the study, it is not possible to indicate how the influence of feature x leads to the choice of *this* in one case, or *that* in another.

3.2 Extraction for an abstract feature representation

Before running the classification process, the data containing the occurrences of *this* and *that* must be extracted so as to present a sequence of features to the classifier. The objective is to convert the previously mentioned characteristics present in texts into an abstract representation composed of lines of features. For each occurrence of the demonstratives, a sequencing PERL program scans the three preceding and following tokens and PoS-tags to match them with the specific features expected to have an impact on the selection of demonstratives in native English. As a result, lines of features are created for each occurrence of the form and a class is assigned to each line of features.

For the first experiment, the program extracts features from the two subsets described in 2. This sequence of features is then matched to a particular class: *expected* or *unexpected*. For all the lines of features extracted from the native subset the class *expected* is assigned. For all the features extracted from the learner subset the class *unexpected* is assigned. Once the classes are assigned, both subsets are merged so as to finalise the training and test sets for the classifier. Illustration 1 is a partial view of an extraction process where linguistic characteristics are turned into lines of features. The second line starts with the feature *of* as it was found three words before an occurrence of *this*, also printed as the second last feature of the same line. The hyphen sign after *of* means that none of the tokens listed in 3.1 were found two words before the occurrence of *this*. When PoS tags are matched by the PERL program they also are printed. The tags PUNCL and NN indicate that some punctuation and a noun were found within three words before the occurrence of *this*. A hyphen denotes a non-existing feature for a given position before or after the occurrence. The last element corresponds to the class assigned. When features are extracted from the native corpus subset described in 2.1, the *expected* class is printed just like line two of the illustration.

For the second experiment, a similar sequencing program is run on the Charliphonia subset described in 2.2 to create lines of features with their assigned class: *this* or *that*. In this experiment, extra features are scanned by the PERL program as the subset is characterised by the fact that it only includes *unexpected* uses of the forms. Since the objective of the experiment is to test features that lead to unexpected use, we add non-native features based on the linguistics characteristics described in 3.1.



```
- - - - - - - PUNCL - NN - - - - - - - MD - PUNCR - - DT this expected
of - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN - TPRON this expected
- - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN - TPRON this expected
of - - - - - - PUNCL - NN - - - - - NN - - - PUNCR IN - TPRON this expected
year - - - - - - PUNCL - NN - (W)RB - - - - - - (W)RB PUNCR - - TPRON this expected
year - - - - - - PUNCL - NN - - - - - - - - - PUNCR - - TPRON this expected
year - - - - - VBD PUNCL - - VB - - - - - - - - PUNCR IN IN TPRON this expected
year - - - - - - PUNCL - - - - - - - - - - - PUNCR IN IN TPRON this expected
- - - - - - - PUNCL - - - - CC - - - - - - PUNCR IN - DT that expected
fall - - - - - - PUNCL - - - (W)RB - - - - - - (W)RB PUNCR - - TPRON this expected
of - - - - - - PUNCL - NN - - - - - NN - - - PUNCR IN - TPRON this expected
of - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN IN TPRON this expected
when - - - - - - PUNCL - - VB - - - - - - - - PUNCR - - DT this unexpected
- - - - - - VBD PUNCL PRP - VB - - - - - - - - PUNCR - - DT this unexpected
like - - - - - - PUNCL - NN - - - - DT/PDT NN - - - PUNCR IN - TPRON this unexpected
like - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN - TPRON this unexpected
like - - - - - - PUNCL - - - - - - DT/PDT - - - - PUNCR IN - TPRON this unexpected
like - - - - - - PUNCL - NN - - - - DT/PDT - - - - PUNCR IN - TPRON this unexpected
n't - - - - - - PUNCL - NN - (W)RB - - - - - - (W)RB PUNCR - - DT this unexpected
- - - - - - - PUNCL - - - - - - - NN - - - PUNCR IN - DT this unexpected
- - - - - - - PUNCL - NN - - - - - NN - - - PUNCR IN - DT this unexpected
- - - - - - - PUNCL - NN - - - - DT/PDT - - - - PUNCR IN - DT this unexpected
- - - - - - - PUNCL - NN - - - - - - VB - - PUNCR IN - TPRON this unexpected
- not - - - - - PUNCL - NN - - CC - DT/PDT - VB - - PUNCR - - TPRON this unexpected
same - - - - - - PUNCL PRP - VB - CC - - - VB - - PUNCR - - TPRON this unexpected
- - - - - - - PUNCL - NN - - - - - - VB - - PUNCR - - TPRON this unexpected
like - - - - - - PUNCL - - - - - - - - VB - - PUNCR - - TPRON this unexpected
- - - - - - - PUNCL - NN - - CC - - - VB - - PUNCR - - TPRON this unexpected
moment - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN - DT this unexpected
that - - - - - VBD PUNCL - - - - - TCOM/TREL - - VB - - PUNCR - - TPRON that unexpected
```

Illustration 1. Feature set for *expected* and *unexpected* classes

4 Classification and results

In this section, we explain the classifying method used by the classifier and how its performance is assessed. The second part deals with the results of the classification experiments.

4.1 Classification method

The machine-learning method used for the experiment applies the memory-based method, and the IB1 or k-nearest neigbour algorithm is implemented in TiMBL (Daelemans et al. 2010). In machine learning, two types of data are necessary in order to classify, and verify the classification and its performance. The memory-based learner TiMBL first goes through a training phase before performing the classifying phase. In the training phase, it adds lines of features and their class to its memory. Each line constitutes a vector of features. In the classifying phase, the classifier predicts the class of new lines of features without the class information. The similarity between the new lines of features and all the examples in memory is computed using some distance metric. The prediction is made by assigning the most frequent category within the found set of most similar line(s), *i.e.*, the k lines memorised in the training phase that are nearest to the line being processed. To do so, the classifier computes a series of metrics (gain ratio) in order to establish the order of the features to be taken into account in the decision process. It establishes a hierarchy of the features from most relevant to least relevant in the classifying process.

Due to the low volume of data, we use the leave-one-out option for training and testing on our dataset, which means that for each instance of the experiment, one line of the file only is used for testing and the other patterns are used for training. This process is repeated for each pattern and the advantage is that, considering the small size of the samples, the leave-one-out option allows for greater robustness and generality to the learned hypothesis in this case. “No test file is read, but testing is done on each pattern of the training file, by treating each pattern of the training file in turn as a test case (and the whole remainder of the file as training cases).” (Daelemans et al. 2010: 41). In order to evaluate the performance of the classification, precision and recall are calculated for each line due to the leave-one-out option. The results presented in 4.2 represent averages of each metric for successive classifying tests.

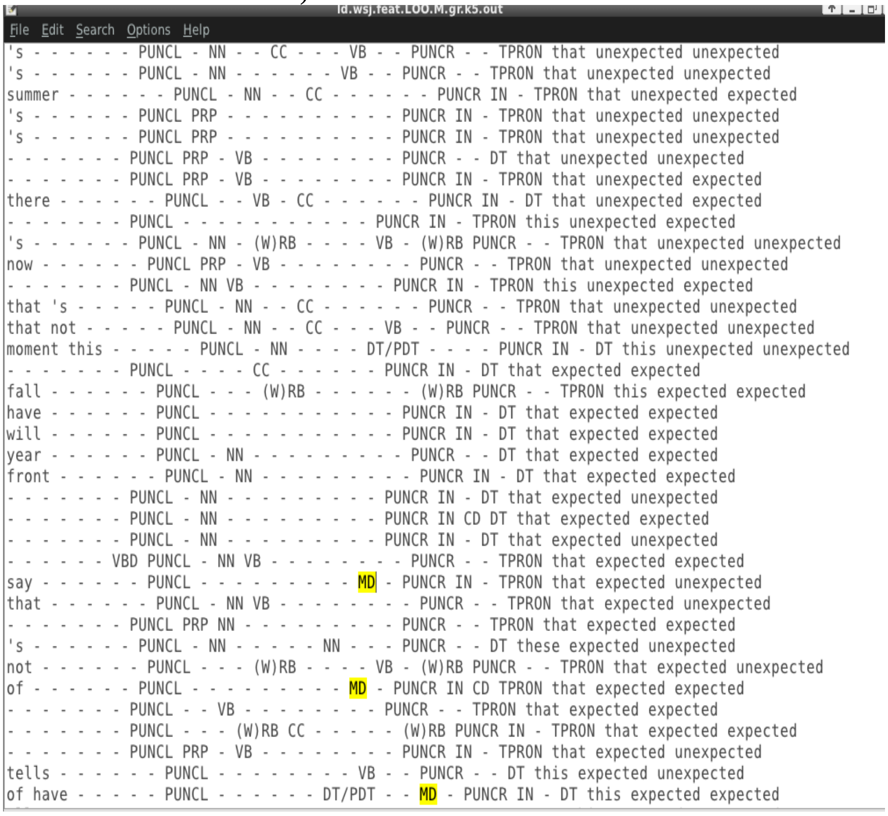
4.2 Results

We present the results of our experiments in two parts. First, we show the results of *unexpected* and *expected* classification in table 2. The subsets used for experiment one include an equal number of *expected* and *unexpected* lines and an equal number of *this* and *that* occurrences. This means that random classification would provide overall accuracy of 50%. Considering this 50/50 baseline of our subsets, the extra 20% improvement margin (the actual accuracy less the random accuracy) gives a measurement of the relevance of the feature set for the selection of expected *this* or *that*. This level remains rather low as NLP classification tasks usually show results well above 90%. The fact that not all lines obtain the correct class may be explained by a lack of exhaustiveness in the type of features. The immediate context of each occurrence, that is three tokens and three PoS-tags before and after may be seen as a limitation as there may be linguistic characteristics located further away that influence the selection of a class. Another limitation may find its source in the difference between the oral and written modes of the native and non-native subsets. The mode of the WSJ is written while that of the non-native subset is oral. A bias may have been introduced due to differences linked to distinct style and syntactic-complexity profiles. Classification between *expected* and *unexpected* uses determines the extent to which the features have an impact on the selection of the forms.

Table 2. Experiment 1 - *Unexpected* and *expected* classification results.

Scores per value class	precision	recall
expected	0.69048	0.72500
unexpected	0.72500	0.69048
overall accuracy: 0.707317		

So, the mixed subset approach shows that it is possible to distinguish between *unexpected* and *expected* forms thanks to the selection of particular features that the classifier uses to categorise the abstraction of occurrences. TiMBL allows the user to have access to the feature order set during the training phase. The gain ratio weight calculated for each feature shows the significance of each feature in the classification. Incidentally, it provides the linguist with significant information on each linguistic characteristic that has been abstracted in the feature vectors. For experiment one, the first four features have a gain ratio above 10%. They are CD (Number) within three PoS tags to the left or right and MD (Modal), TCOM/TREL (*that* as complementizer or relative pronoun) and DT/PDT (Determiner or pre-determiner) within 3 PoS tags to the left. For the classifier, the presence of a cardinal number in the left or right context is a prime criterion to differentiate between *expected* and *unexpected* uses of any demonstrative by learners. If we look at the data, it appears that CD is only used in the *expected* subset. So, it is logical that any new line including the CD feature is classified as *expected*. The TCOM/TREL feature only appears once in the data and so it also becomes a determining factor for classification in logical terms. It may have been more appropriate to search for this tag in the right context of the forms as it can be argued that hypotaxis occurs to provide details of an entity expressed in its preceding NP. This search may have led to more occurrences of this feature, giving it more relevance in linguistic terms. So the relevance of the CD and TCOM/TREL features may be questioned linguistically as their higher gain ratio may only be due to the data representation of the sample. On the contrary, the distribution of DT/PDT shows a different pattern in the data as it is found in many lines, but not all, for both classes. Classification shows that when the feature appears on a line, it leads to *unexpected* in 9 cases out of 11. So, while this feature is present across the data, the classification results suggest that it plays a significant role in helping the classifier differentiate an *expected* use of a demonstrative from an *unexpected* use. Data suggest it might be a feature of unexpectedness. The MD feature and its influence remains unclear. In the training data, it appears in *expected* uses only so it is logically found in lines classified as *expected*. However, it appears on one line classified as *unexpected* (For all comments see Illustration 2).



```
's - - - - - - PUNCL - NN - - CC - - - VB - - PUNCR - - TPRON that unexpected unexpected
's - - - - - - PUNCL - NN - - - - - - VB - - PUNCR - - TPRON that unexpected unexpected
summer - - - - - - PUNCL - NN - - CC - - - - - - PUNCR IN - TPRON that unexpected expected
's - - - - - - PUNCL PRP - - - - - - - - - - PUNCR IN - TPRON that unexpected unexpected
's - - - - - - PUNCL PRP - - - - - - - - - - PUNCR IN - TPRON that unexpected unexpected
- - - - - - - PUNCL PRP - VB - - - - - - - - PUNCR - - DT that unexpected unexpected
- - - - - - - PUNCL PRP - VB - - - - - - - - PUNCR IN - TPRON that unexpected expected
there - - - - - - PUNCL - - VB - CC - - - - - - PUNCR IN - DT that unexpected expected
- - - - - - - PUNCL - - - - - - - - - - - PUNCR IN - TPRON this unexpected expected
's - - - - - - PUNCL - NN - (W)RB - - - - VB - (W)RB PUNCR - - TPRON that unexpected unexpected
now - - - - - - PUNCL PRP - VB - - - - - - - - PUNCR - - TPRON that unexpected unexpected
- - - - - - - PUNCL - NN VB - - - - - - - - PUNCR IN - TPRON this unexpected expected
that 's - - - - - PUNCL - NN - - CC - - - - - - PUNCR - - TPRON that unexpected unexpected
that not - - - - - PUNCL - NN - - CC - - - VB - - PUNCR - - TPRON that unexpected unexpected
moment this - - - - - PUNCL - NN - - - - DT/PDT - - - - PUNCR IN - DT this unexpected unexpected
- - - - - - - PUNCL - - - - CC - - - - - - PUNCR IN - DT that expected expected
fall - - - - - - PUNCL - - - (W)RB - - - - - - (W)RB PUNCR - - TPRON this expected expected
have - - - - - - PUNCL - - - - - - - - - - - - PUNCR IN - DT that expected expected
will - - - - - - PUNCL - - - - - - - - - - - - PUNCR IN - DT that expected expected
year - - - - - - PUNCL - NN - - - - - - - - - - PUNCR - - DT that expected expected
front - - - - - - PUNCL - NN - - - - - - - - - - PUNCR IN - DT that expected expected
- - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN - DT that expected unexpected
- - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN CD DT that expected expected
- - - - - - - PUNCL - NN - - - - - - - - - PUNCR IN - DT that expected unexpected
- - - - - - VBD PUNCL - NN VB - - - - - - - - PUNCR - - TPRON that expected expected
say - - - - - - PUNCL - - - - - - - - - MD - PUNCR IN - TPRON that expected unexpected
that - - - - - - PUNCL - NN VB - - - - - - - - PUNCR - - TPRON that expected unexpected
- - - - - - - PUNCL PRP NN - - - - - - - - - PUNCR - - TPRON that expected expected
's - - - - - - PUNCL - NN - - - - - NN - - - PUNCR - - DT these expected unexpected
not - - - - - - PUNCL - - - (W)RB - - - - VB - (W)RB PUNCR - - TPRON that expected unexpected
of - - - - - - PUNCL - - - - - - - - - MD - PUNCR IN CD TPRON that expected expected
- - - - - - - PUNCL - - VB - - - - - - - - PUNCR - - TPRON that expected expected
- - - - - - - PUNCL - - - (W)RB CC - - - - - (W)RB PUNCR IN - TPRON that expected expected
- - - - - - - PUNCL PRP - VB - - - - - - - - PUNCR IN - TPRON that expected unexpected
tells - - - - - - PUNCL - - - - - - - - VB - - PUNCR - - DT this expected unexpected
of have - - - - - PUNCL - - - - - - DT/PDT - - MD - PUNCR IN - DT this expected expected
```

Illustration 2. Lines of features with their initial and automatically assigned classes.

Experiment two gives information on learner specific features that lead to *unexpected* uses. This is why the second experiment is based solely on *unexpected* uses of learners, as it is expected to give more insight into the selection process of *unexpected* forms. The following results (table 3) are in relation to the classification of *unexpected* forms only. There are two phases.

Firstly, the experiment is carried out with features only selected for native English, or in other terms, the features used in the first experiment. Secondly, the learner features mentioned above are added. For example, if we consider the column “overall accuracy”, we obtain 0.80 accuracy when the features are extracted with non-learner specific features, and 0.88 when they are extracted with learner-specific features. The gain in accuracy is real, and shows that these features have an impact on the selection of *unexpected* forms. Even if more features related to learner use need to be tested, the process shows that it is possible to validate learner-related features that lead to *unexpected* uses. It also shows that features based on native English also partake in the *unexpected* form selection.

If we study the order of the feature weights calculated by the classifier with non-learner specific features, the following features appear first: TCOM/TREL (*that* complementizer or relative pronoun), IN (Preposition), CC (Coordinating conjunction). The same calculation with learner-specific features gives the following order: NNS (plural noun), TCOM/TREL *(that* complementizer, relative pronoun), IN (Preposition) and CC (Coordinating conjunction). So, the way to distinguish learner unexpected uses of *this* from uses of *that* is done primarily *via* a feature denoting plural agreement error and it has a significant impact on overall accuracy. When observing the data, it appears that NNS is always linked to the unexpected selection of *this*. The following example shows the word *countries* in context:

(4) Speaker A: I haven't class this day er
Speaker B: sports
Speaker A: em not no sports but em music em because they they em in in *this countries* em er em movement er musical movement was born in the nineteenth er twentieth century

The fact that *countries* was POS tagged as a plural noun with NNS, and that NNS was passed on to the classifier as a feature, made the classifier learn that the sequence *this* + NNS was not possible. As a consequence, any new occurrence of the sequence in any context would be classified as *unexpected*.

The second group of features remains the same as with the non-learner specific extraction. In 9 cases out of 10, CC is related to the correctly assigned *that* class, making it a candidate for influencing the unexpected selection of *that*. Conversely, IN corresponds to 11 cases of correctly assigned *this* as opposed to 4 cases of correctly assigned *that,* which would make it a candidate for influencing the unexpected selection of *this*. To finish, TCOM/TREL appears only once in the data, which makes it a logical but not linguistically relevant factor.

Table 3. Experiment 2 – Classification of *unexpected this* and *that*, with and without learner specific features.

Scores value class:	overall accuracy		precision		recall	
	Non learner specific	Learner specific features	Non learner specific	Learner specific features	Non learner specific	Learner specific features
this	N/A	N/A	0.77273	0.85714	0.85000	0.90000
that	N/A	N/A	0.85000	0.90476	0.77273	0.86364
	0.809524	0.880952				

5 Conclusion

In this article, we have covered the issue of automatic classification of learner English occurrences of unexpected uses of *this* and *that*. We have evaluated two types of classifications based on native and learner English. The first objective was to see how *unexpected* learner use of *this* and *that* could be predicted among occurrences including native and non-native use of the forms. The second objective was to uncover elements that may influence the selection of *unexpected* forms by learners. Answers to this second question would provide valuable information to support teaching to ESL students as teachers could make their students aware of specific recurrent features of unexpected uses of the forms. The approach adopted in this study has allowed the contrastive exploration of non-native speech and native speech with the aim of finding features that influence linguistic choices. In line with Frei's view, errors are used as traces of language needs to be explained rather than to be compared with a norm.

Two types of data were used. In the first experiment, the classifier was trained with two equal subsets composed of native and non-native corpora. The non-native subset consisted of learner uses of *this* and *that* in an *unexpected* manner. Special care was given to the selection of features so as to make them correspond to linguistic notions developed in the literature on *this* and *that*. The classifying process was a test to distinguish between *expected* and *unexpected* forms, and results showed a 70% accuracy. The presence of determiners or predeterminers in the immediate context appeared as a significant feature for unexpectedness.

The second experiment was carried out to have closer insight into the selection of *unexpected* forms by learners. Several features were tested to measure the extent of their importance in the selection process. The feature related to plural nouns, i.e., a plural noun after a singular form of *this* or *that*, proved to have a substantial impact on classification as it improves the performance by 7% compared with classification based on features identified on native English. Overall, features such as *plural noun* and *preposition* in the previous context seem to be factors for unexpected *this.* The *coordinating conjunction* feature may be a factor for the unexpected selection of *that.*

Future work includes the refinement of the feature selection process as the one based on native English still needs more accurate and relevant information. For example, each occurrence of *this* and *that* could be annotated as being deictic or anaphoric. This kind of feature might help to identify contexts and their type of reference. More learner specific features also need to be identified and tested to explore further the way learners operate their choices while speaking. These features also need to be tested on other non-native corpora including NOCE, a spoken corpus of learner English (Diaz Negrillo 2009). After identifying features of use for the demonstratives both in native and non-native speech, it will then be possible to automatically introduce them in an annotation layer. As a result, the full process of detection and annotation will be automated. Queries on the demonstratives may then be launched on several corpora simultaneously. The ultimate objective will be to carry out comparative analysis on the use of demonstratives, between learners of different L1s, or between natives and learners. This classification method could also be applied to other linguistic items. The process would require the selection of linguistic contextual features for the given item. Entire corpora could hence be processed to classify all occurrences of the item according to relevant categories for the analysis of this item.

Learner English is a fast-growing field of study and has been accompanied by the development of many corpora. Each corpus comes with its own meta-structure which makes cross-corpora querying impossible. By focusing on a particular linguistic point, our project is to develop a fine-grained automatic annotation process that will be applicable to any corpus. The final objective is to make it possible to import and query several corpora at the same time in order to carry out contrastive analyses depending on the nature of these corpora (native v. non-native, different L1s).

References

Biber, D., S. Johanson, G. Leech, S. Conrad and E. Finegan (1999), *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Charniak, E., D. Blaheta, N. Ge, K. Hall, J. Hale and M. Johnson (1987), 'BLLIP 1987-89 WSJ Corpus Release 1'

Cornish, F. (1999), *Anaphora, Discourse, and Understanding. Evidence from English and French*. Oxford: Oxford University Press.

Daelemans, W., J. Zavrel, K. van der Sloot and A. van den Bosch. (2010), *TiMBL: Tilburg Memory-Based Learner Version 6.3 Reference Guide*. Tilburg, The Netherlands: Induction of Linguistic Knowledge, Tilburg University and CLiPS, University of Antwerp, Available online at http://ilk.uvt.nl/downloads/pub/papers/ilk.1001.pdf (last accessed on June 17, 2013).

Dagneaux, E., S. Denness and S. Granger (1998), 'Computer-aided Error Analysis' *System*, (26): 163–174.

Díaz Negrillo, A. (2009), *EARS: A user's manual.* Munich. LINCOM Academic Reference Books.

Fraser, T., A. Joly (1979), 'Le système de la deixis - Esquisse d'une théorie d'expression en anglais' *Modèles linguistiques*, 1: 97–157.

Frei, H. [1929] (2011), *La Grammaire des fautes*. Rennes: Presses Universitaires de Rennes.

Gaillat, T. (2013a), 'Towards a fine-grained annotation of *this* and *that*: a typology of use in native and learner English' in: S. Granger, G. Gilquin and F. Meunier (eds). Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use - Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S. (2008), 'Learner Corpora in Foreign Language Education' In *Encyclopedia of Language and Education*, 4: 337–351.

de Haan, P. (2000), 'Tagging Non-native English with the TOSCA-ICLE Tagger' In: C. Mair, M. Hundt (Eds.) *Corpus Linguistics And Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora*, Amsterdam & Atlanta GA: Rodopi B.V., 69–80

Halliday, M. and R. Hasan (1976), *Cohesion in English*. English Language Series. Harlow: Pearson Education Limited.

Han, N. R., M. Chodorow and C. Leacock (2006), 'Detecting Errors in English Article Usage by Non-native Speakers' *Natural Language Engineering,* 2 (12): 115–129.

Levy, R. and G. Andrew (2006), 'Tregex and Tsurgeon: Tools for Querying and Manipulating Tree Data Structures' *Proceedings of the 5th International Conference on Language Resources and Evaluation,* Available online athttp://nlp.stanford.edu/pubs/levy_andrew_lrec2006.pdf (last accessed on June 17, 2013)

Marcus, M. P., M. A. Marcinkiewicz and B. Santorini (1993), ‘Building a Large Annotated Corpus of English: The Penn Treebank’ *Computational Linguistics*, 19. 313–330.

de Mönnink, I. (2000), ‘Parsing a Learner Corpus’ In *Corpus Linguistics and Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora*, 82–90. Amsterdam & Atlanta GA: Rodopi B.V.

Pradhan, A. M., A. S. Varde, J. Peng, and E. M. Fitzpatrick (2010), ‘Automatic Classification of Article Errors in L2 Written English’ In *Proceedings of the Twenty-Third International FLAIRS Conference*. Association for the Advancement of Artificial Intelligence (AAAI), Available online at https://www.aaai.org/ocs/index.php/FLAIRS/2010/paper/view/1342/1751, (last accessed on June 17, 2013).

van Rooy, B. and L. Schafer (2003). ‘An Evaluation of Three PoS Taggers for the Tagging of the Tswana Learner English Corpus’ In *Proceedings of the Corpus Linguistics 2003 Conference*, Available online at http://www.corpus4u.org/upload/forum/2005092023174960.pdf (last accessed on June 17, 2013)

Schmid, H. (1994), ‘Probabilistic Part-of-Speech Tagging Using Decision Trees’ In *Proceedings of the International Conference on New Methods in Language Processing.* Available online at http://www.stttelkom.ac.id/staf/imd/Riset/POS%20Tagging/Using %20Decision%20Tree.pdf (last accessed on June 17, 2013)

Stirling, L. (2002), ‘Deixis and Anaphora’, in: R. Huddleston and G. K. Pullum (eds) *The Cambridge Grammar of the English Language*, Cambridge University Press. Beccles, Suffolk, 1449–1564