A Multifactorial analysis of *this*, *that* and *it* proforms in anaphoric constructions in learner English

Analyse multifactorielle des proformes this, that et it dans des constructions anaphoriques chez des apprenants d'anglais

Thomas Gaillat

Electronic version

URL: <http://journals.openedition.org/praxematique/5668>
DOI: 10.4000/praxematique.5668
ISSN: 2111-5044

Publisher

Presses universitaires de la Méditerranée

Electronic reference

Thomas Gaillat, « A Multifactorial analysis of *this*, *that* and *it* proforms in anaphoric constructions in learner English », *Cahiers de praxématique* [Online], 72 | 2019, Online since 26 June 2019, connection on 08 September 2020. URL : <http://journals.openedition.org/praxematique/5668> ; DOI : <https://doi.org/10.4000/praxematique.5668>

This text was automatically generated on 8 September 2020.

Tous droits réservés

A Multifactorial analysis of *this*, *that* and *it* proforms in anaphoric constructions in learner English

Analyse multifactorielle des proformes this, that et it dans des constructions anaphoriques chez des apprenants d'anglais

Thomas Gaillat

The author would like to thank Andrew Simpkin, Lecturer in Statistics at the School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, for his advice on the modeling process.

Introduction

- 1 Learner Corpus Research is a domain in which the morphosyntactic and semantic dimensions of language have been studied extensively (Granger, 1994; Barlow, 2005; Gries, 2008; Díaz-Negrillo *et al*, 2013; Granger *et al*, 2015) with a view to discovering developmental patterns in interlanguage (Selinker, 1972). However, the pragmatic dimension has not benefited from the same level of attraction even though learners do experience difficulties of a discursive nature. Anaphora falls in this category, which includes the use of *it*, *this* and *that* in learner English. Several studies have brought evidence of idiosyncrasies in the use of the forms (Petch-Tyson, 2000; Lenko-Szymanska, 2004; Gaillat, 2013; Zhang, 2015), highlighting the existence of a learner-specific proform microsystem (Gaillat, 2016). Understanding these idiosyncrasies would give insights in how learners build anaphoric processes.
- 2 The difficulties experienced by learners to select between *it*, *this* and *that* seem to find their origins in the way learners understand anaphoric features (Gaillat, 2016). They have a partial representation of the anaphoric microsystem as they do not integrate the full spectrum of linguistic criteria used for the selection of a form. This erroneous representation generates confusion due to a competition between the three forms on

the paradigmatic axis. The source of such confusion is difficult to determine due to the multidimensional representation of the criteria. In order to understand the confusion, it is necessary to analyse the representations that learners have of the microsystem. This can be achieved by comparing their productions with those of natives.

- 3 Current studies have focused on analysing the forms in two manners. Some studies conduct experiments based on one corpus. They use one learner corpus and analyse the proforms internally. Some studies only take one or two of the three forms into account, hence ignoring interactions within the microsystem. These approaches have yielded significant results but hinder the comparative dimension of the three forms across different corpora representing various L1s. Our approach is based on two corpora of different L1s (Spanish and Native English) to support a Contrastive Interlanguage Analysis (CIA) (Granger, 1996). As there are interactions within the microsystem, we analyse the occurrences of all three proforms within a multi-factorial framework. We conduct a logistic regression analysis to identify the factors that significantly influence their selection. This paper is divided into five sections. Section 1 covers work related to anaphora and learner English. Section 2 provides details about the annotated corpora we used in our experiment. In Section 3, we detail the experimental set up and the results are presented in Section 4. We discuss the results and draw perspectives in Section 5.

1. Related Work

- 4 Our approach is grounded in Discourse Functional Grammar (Cornish, 1999). Within a space of potential discourse referents, a speaker orientates the addressee towards the relevant target (i.e. the referent) by selecting anaphoric expressions such as pronouns and determiners. As anaphoric expressions, *it*, *this* and *that* have a set of referential characteristics which encode their referential potential. Their selection results from a process of adequating their referential potential to the characteristics of their context such as the degree of givenness of the referent (Gundel *et al*, 1993), their level of accessibility (Ariel, 1988; Kleiber, 1992), of focus (Strauss, 2002) and also their scope (Lapaire and Rotgé, 1991: 50). This selection also depends on the morphosyntactic characteristics of the cotext (Cornish, 1999: 82), as well as the pragmatic contextual characteristics of endophora and exophora (Halliday and Hasan, 1976). All of these characteristics form a cognitive representation of anaphoric features. This representation enables the speaker and the addressee to narrow down the space of potential referents in the process of anaphora resolution (Scott, 2013).
- 5 *It*, *this* and *that* have been the subject of several studies based on quantitative linguistics methods. Some researchers looked at one form in particular (Wulff *et al.*, 2012) in one corpus. Others took the three forms into consideration but relied on one native corpus (Strauss, 2002; Wang *et al*, 2011). In order to analyse learner use of the forms, some studies followed a Contrastive Interlanguage Analysis (CIA) in order to compare Native Speakers (NS) and Non-Native Speakers (NNS) (Granger, 1996). However researchers, in these cases, only analysed *this* and *that* as a binary microsystem (Petch-Tyson, 2000; Lenko-Szymanska, 2004; Liang, 2009). To the best of our knowledge only one study relied on a combination of several corpora (learner and native) for the study of the three forms (Zhang, 2015). A frequency analysis of occurrences was complemented by a

qualitative analysis of influencing factors. However, the effects of the factors were not measured.

- 6 All of the aforementioned studies implemented statistical methods based on frequencies of occurrences of the forms together with significance indicators, which does not provide information on the features that influence the selection of any of the forms. None of the studies included a quantitative multifactorial approach in which different contextual features were considered as a combination of factors influencing the use of any of the three forms. This type of experiment is needed as it would address the question of identifying and quantifying the effects of the factors that are correlated to the forms. Our approach tries to fill this gap by addressing both the NS/NNS contrast and the factor identification requirement. We employ a multi-factorial analysis in order to take advantage of the explanatory power of the variables much in the same way as studies on article placement and dative alternation (Bresnan et al., 2007; Gries, 2003, 2013). This type of approach deals with the multidimensional complexity of microsystems whose outcome variables, i.e. the selected forms, depend on a number of factors which need to be explored.

2. Corpora and Annotation

- 7 In this section we describe the corpora we used and the annotation scheme that was applied to the texts.

2.1. Corpora

- 8 In order to integrate a CIA perspective in our approach, we make use of two written English corpora of two different first languages (L1). We use the Wall Street Journal (WSJ) module of the Penn Treebank project (Marcus *et al.*, 1993), a native American English corpus of circa one million words from news articles. The second corpus, called NOCE (Díaz-Negrillo, 2007), is made up of argumentative texts written by Spanish learners of English. As some of the annotations had to be manually added, we used a subset of each corpus. The WSJ subset is made up of 40,726 tokens in 96 independent articles in total, and the NOCE subset includes 14,446 tokens in 46 texts from distinct learners.
- 9 The two corpora represent written English. They both include texts of argumentative speech which, in this respect, makes them comparable. However, they differ in terms of topics. While the learners worked on the same task, i.e. giving their opinion on teaching religion in the Spanish school system or telling a past experience, the journalists of the WSJ reported news on various topics. In their articles they supported their opinions with descriptive factual news reports. This difference may induce different trends in the use of anaphoric expressions. For instance, many references to dates were used in the WSJ articles, while very few were collected in the NOCE corpus.

2.2. Annotation Scheme

- 10 Prior to building the dataset the two subsets were annotated semi-automatically by applying the same annotation scheme. We applied manual annotation for pragmatic features and automatic annotation for morphosyntactic features.

- 11 One annotator (the author) carried out the annotation of the pragmatic features in both corpora. Prior to annotating, annotation rules were defined to clearly determine the linguistic context and structures of each label. Table 1 summarises the two types of tags and the context in which they appear. Two types of labels were added to each occurrence of a proform. Firstly, endophoric and exophoric labels were added to reflect the situational or textual location of the referent (Halliday and Hasan, 1976). We called it the context annotation layer. Secondly, the accessibility status of each referent was also added manually to each referring expression. We called it the discourse annotation layer. It is encoded by providing different degrees of givenness of the referent (Gundel *et al*, 1993) as implemented by Komen (2013). The tags indicate different degrees of accessibility, i.e. identity between referent and anaphor, existing focus on referent in situation of utterance, referent inference, new referent or non-existent referent.

Table 1: Manual annotation scheme applied to the two corpus subsets

Context tags	Discourse tags	Description (from the addressee's point of view)
EXO	NEW	Referent in the situation of utterance. Unfocused so far. New information
ENDO	NEW	No antecedent given so far up to the point of mention in the discourse but the referent is in the discourse. New information.
EXO	ONFOC	Some degree of focus on referent (implicitly or explicitly in the addressee's point of view) in the situation. Already given information.
ENDO	INFRD	The anaphor has an antecedent in the text, but the referents of the current anaphor and its antecedent are not the same (they can be in a part-whole relation or reference may originate from a clause rather than a single word). The mention of the first noun phrase must already have implied the existence of the second noun phrase, which infers from it, i.e. associative anaphora linking several entities. Already given information.
ENDO	IDTY	The anaphor has an antecedent in the text, and the referents of both are identical. Already given information
NA	INRT	The item does not have an antecedent inside or outside the text, and it cannot be referred to in the following context.

- 12 Concerning morphosyntactic features, we used several automated methods to apply annotation. Part-of-Speech (POS) tags were applied with TreeTagger (Schmid, 1994) by using a modified version of the Penn Treebank POS tagset (Marcus *et al*, 1993). This version includes distinctions for *this* and *that*, depending on their determiner and proform functions among others such as complementizer and relativiser for *that*. It also includes a distinction between non-referential and pronoun *it*. TreeTagger relies on a statistical model which was trained on the modified tagset. Its evaluation on a test set of the WSJ showed 96.55 % F-Score for *this* proform, 90.19 % F-Score for *that* proform and 90.08 % for *it* (see Gaillat, 2016 Section 5.1.1 for details).

- 13 The nominative case of each occurrence of any of the three forms was also added in order to capture the importance of their position as subject of a predicate or not (nominative vs oblique). This was achieved with a PERL program¹ in order to distinguish the subject position from any post-verbal position. To do so, we applied a rule based approach which relies on patterns of POS tags. For instance, should the token *it* as a pronoun be found in front of a verb, it would automatically be tagged as a nominative case. All occurrences of each proform were tagged with the NOMI and OBLI labels. The tagging algorithm was tested in previous work on the same corpora. Results showed 0.94 F-Score for the NOMI label and 0.87 F-Score for OBLI (see Gaillat, 2016 Section 5.1.2 for details).
- 14 As a result of the manual and automatic annotation processes, all occurrences were tagged with different labels resulting in several annotation layers. The following examples from the Penn Treebank show how some utterances are annotated. In example [1] *that*, as a proform (TPRON) and subject (NOMI) of BE, is used to refer to a previously mentioned (ENDOpheoric) entity in the text. This discourse entity is constructed by associating several concepts, i.e. person, work, speedometer. *That* is used to refer to their combination.

[1] If a new person got to work on part of the speedometer, *that* (TPRON, ENDO, NOMI, INFRD) was a big deal.

- 15 In example [2], there are two annotated forms. *This*, as a determiner (DT) in a non-subject Noun Phrase (OBLI), is used to refer to a new entity in the discourse, i.e. fall. To be able to locate the season, the addressee requires the year which is situational knowledge (EXOpheoric). *It*, as a personal pronoun (PRP) in a non-subject position, is used to endopheorically (ENDO) refer to the value of the dollar being driven down, a specifically identified entity (IDTY) in the discourse and the cotext (also called the antecedent).

[2] Treasury Undersecretary David Mulford defended the Treasury's efforts *this* (DT, EXO, OBLI, NEW) fall to drive down the value of the dollar, saying *it* (PRP, ENDO, NOMI, IDTY) helped minimize damage from the 190-point drop in the stock market Oct.13.

3. Experimental setup: Dataset and Modelling Method

- 16 A corpus can be considered as a set of observations of many linguistic items. To apply a modelling approach on a corpus means that specific items need to be observed in relation to others. A statistical model uses the corpus items as variables. It corresponds to a mathematical function that relates predictor variables to one outcome variable also called response. The purpose is to find a function that can explain most of the observations. To do so, it uses data to “learn” the closest formula that can predict an observable outcome. In our experiment, we want to relate the observable choice of a proform to a set of other predictor variables. Prior to “learning”, modelling requires the use of a structured set of variables. Consequently, the textual and annotation items of a corpus must be transformed into a structured set of values categorised into variables.

3.1. Dataset and Variables

- 17 To build the dataset², all occurrences of the three proforms (by using their POS) were extracted from the corpora and placed in a table-like structure in which each variable value pair corresponds to a text or annotation item. Table 2 summarises the variables, their possible values and their linguistic description. As each occurrence was used as an observation, other features sourced from the annotation of this occurrence were also collected. As a result each observation in the dataset corresponds to a line including one proform and its linguistic features which are subsequently treated as variables in the modelling process.

Table 2: Variable value pairs used in the dataset

Variables	Values	Description
featVBZ	VBZ or ‘-’	The presence of a verb in the present simple tense in the 3-gram context of the form
featED	ED or ‘-’	The presence of a verb in the past simple tense in the 3-gram context of the form.
featNOT	NOT or ‘-’	The presence of negation in the 3-gram context of the form.
featPUNC	PUNC or ‘-’	The presence of strong punctuation in the close context of the form.
featREFPRON	REFPR or ‘-’	The presence of a referential personal or demonstrative pronoun in the close context.
Context	ENDO or EXO	The endophoric or exophoric nature of the form.
Discourse	INFRD, IDTY, NEW, INRT or ONFOC	The givenness nature of the form
tagsCASE	NOMI or OBLI	The nominative case or not of the for
Corpus	NOCE or WSJ	The corpus is used as a variable to indicate the L1 of the speakers.

- 18 The variables intend to represent syntactic, semantic and pragmatic features of the contexts in which the forms appear. Tense (i.e. present simple or past simple) semantically shows whether the speaker introduces a temporal distance between the moment of utterance and the entity referred to (Biber et al., 1999, 347). Negation shows a possible rejection by the speaker impacting the distinction between *this* or *that* (Fraser and Joly, 1979). In terms of syntax, punctuation shows whether the form appears at the beginning or the end of a sentence. This position informs on the potential topicalisation of the form. Likewise, the nominative case intends to draw on the role of syntactic structure in bringing a referent into focus (Gundel *et al*, 1993). The presence of a referential pronoun in the close context informs of the syntactic position

of the form in relation to other elements of a possible referential chain. The dataset includes two variables that provide pragmatic information. The degree of givenness indicates whether the referent is known or not, with various degrees as explained in Section 2.2. The exophoric or endophoric value of the form shows whether the referent is to be identified in the situation of utterance or within the actual context.

- 19 The representation of the dataset follows that presented in Table 3 which includes three examples of occurrences of *that*. For each token (last column) a number of variables (Line 1) take on different values (Lines 2 to 4). For instance, *that*, on line 2, is associated with the NOMI value of the *tagsCASE* variable as well as other values of the same line. On line 4 it is associated with the OBLI value.

Table 3: Extract of the representation of three occurrences of the forms in the dataset

id	Corpus	tags CASE	context	discourse	feat VBZ	feat ED	feat NOT	feat PUNC	feat REFPRON	tokens
GR-2-A-046-10-21	NOCE	NOMI	ENDO	IDTY	VBZ	-	NOT	-	-	that
GR-2-A-046-10-78	NOCE	NOMI	ENDO	INFRD	VBZ	-	-	-	-	that
GR-2-A-046-10-111	NOCE	OBLI	ENDO	IDTY	-	-	-	-	REFPRON	that

- 20 In terms of occurrences, the dataset includes an unequal number of forms as presented in Table 4.

Table 4: Number of occurrences of each form in the dataset

Tokens	It	This	That
All observations	311	66	51

3.2. Modelling Method

- 21 Once the data is formatted, the next step is to choose a type of model (a function) that can adapt (be “fitted”) to the data. Many methods exist and depend mainly on the type of data that are in the dataset (categorical, continuous or ordinal). The choice of the model also depends on whether variable values vary randomly or are fixed. Once the model type is chosen, all variables are included in it and evaluation tests are conducted to determine its power with indicators such as R^2 . By adopting a stepwise selection of the dataset variables, it is possible to eliminate non-significant variables that impair the overall explanatory power of the model. Once it is stable variable effects can be analysed.
- 22 Due to the fact that our dataset is skewed in terms of number of occurrences per form, and because there are many occurrences per individual, it is important to choose a modelling method that takes this type of variation into consideration. We employ a mixed-effect modelling approach (Gries, 2015) as it provides for random effects and fixed effects variables. Random effects account for per-subject variability, i.e. the fact that the dataset does not include the same number of occurrences per speaker. The approach is well suited for the categorical type of the data.

- 23 The model represents a proform as a function of several variables that are assigned weights. As R's implementation (R Core Team, 2012) of mixed-effect modelling with categorical data can only take two outcome variables, we decided to conduct three pairwise analyses of the occurrences of the forms. In other terms three binomial models (Generalized linear mixed model with maximum likelihood (Laplace Approximation)) are fitted to the data. We adopt a stepwise model selection procedure based on ANOVA tests and highest AIC values (Burnham and Anderson, 2004).
- 24 The procedure includes the following steps: after selecting the model, we provide the model summary, the p-values of the variables and the effects of predictor variables on the response variable (i.e. the proform). We also evaluate the model in terms of classification accuracy and explanatory power with the R^2 indicator (Gries, 2015).

4. Results

- 25 We give the results obtained for each mixed-effects model. Each model was computed based on the dataset combining the NOCE and WSJ data.

4.1. *It/this* mixed-effects model

- 26 The model selection procedure leads to the removal of a number of variables as ANOVA shows no significant difference with more complex models. The ANOVA test suggests that tense, referential pronouns and negation variables do not significantly add to the model's explanatory power and may be dropped. AIC³ (275 down from 311.3 with the initial model) also suggests that only *corpus*, *context*, *discourse* and *tagsCASE* may be included as fixed variables.
- 27 The mixed-effect binomial model (see Table A.1 in the Appendix) for *it* and *this* indicates (see Figure 1) that *this* is less likely (under 50 %) to occur than *it* in all cases. However, when the referent is inferred (the *Inferred* variable is very significant with $p < 0.01$) or new (the *new* variable shows $p < 0.01$), to a lesser extent, the chances of *this* increase. For instance the probability of *this* is 40 % when the referent is inferred. The pink segment shows the 95 % Confidence Interval of the estimate.
- 28 The model also shows that the *TagsCASE* variable-value pair favours *this* (see Figure 2), indicating a slight preference for post verbal positions regardless of the language. The probability of *this* is above 10 % in the case of OBLI as opposed to less than 10 % with NOMI. Figure 3 gives an insight in the type of L1 that influences most this observation. The *corpus* and *tagsCASE* variables combine together to produce a significant effect, i.e. the interaction between the *NOCE* and the *OBLI* values is significant ($p = 0.024 < 0.05$). The combination indicates a preference for *this* OBLI in the Spanish L1 corpus. The right panel of the diagram shows that, when the combination *NOCE* and *OBLI* occurs, the chances of getting *this* are higher than if combining the *WSJ* with *OBLI*. It means that Spanish speakers are likely to use *this* in nearly 20 % of the cases when native speakers are likely to use *this* in 10 % of the cases. This shows that Spanish speakers have a slight tendency to overuse *this* in object position. However, the confidence intervals (in pink) weaken the observation.

Figure 1: The effects of the *discourse* variable values on the selection of *this* compared with *it* Figure 2: The effects of the tagsCASE variable values on the selection of *this* compared with *it*

Figure 3: The effects of the *corpus:tagsCASE* variable interaction on the selection of *this* compared with *it*

The explanatory power of the model was computed with marginal $R^2=0.333$ and conditional $R^2=0.612$ ($R=1$ would indicate perfect explanatory power). In terms of classification accuracy the model predictions were matched with actual outcome values in the data. Results show 0.9 accuracy. To compare we computed predictions on the basis of the proportions of each form in the data (a baseline). We obtained a 0.82 accuracy. The model shows a slightly better classification power.

4.2 *It/that* mixed-effects model

- 29 Model selection suggests dropping the tense, referential pronouns, punctuation and negation features as there is no significant difference with the model that includes them. The final model (AIC =234.1 down from 268.1) corroborates the ANOVA. As a result the final model includes the following variables: *speakerID*, *corpus*, *context*, *discourse* and *tagsCASE*.
- 30 The *it/that* model (see Table A.2 in the Appendix) shows that, when the referent is inferred (*discourseINFRD* variable-value pair with $p<0.001$) or new in the discourse (*discourseINFRD* variable-value pair with $p<0.01$), *that* is more likely to occur than in other discourse states. Figure 4 shows an example of visualisation of such preferences. The line shows the probability of choosing *that* as opposed to *it*, and as the degree of givenness changes, the chances of selecting *that* vary. For instance, the chances of *that* are close to 40 % in the case of an inferred referent and 20 % for a new referent. There is no significant effect linked to the type of corpus which indicates a similar use between learners and natives in this case.

Figure 4: The effects of the *discourse* variable on the selection of *that*

The model shows marginal $R^2 = 0.64$ and conditional $R^2 = 0.746$, which indicates a robust explanatory power. Classification accuracy is 0.91 compared with 0.86 in the case of simple classification based on the proportions of each form frequencies.

4.3 *This/that* mixed-effects model

- 31 Model selection shows that the initial model can be simplified as the tense, negation, pronoun, and punctuation variables do not significantly enhance it. The selected model shows a 161.8 AIC compared with 163.9 for the initial model. It includes *speakerID*, *corpus*, *context*, *discourse* and *tagsCASE* as variables.
- 32 The *this/that* model (see Table A.3 in the Appendix) shows several significant variables, i.e. the corpus type, the exophoric context type of reference, the inferred and new discourse types and the oblique case type. The plotted effects are shown in Figures 5 to 8 and show the likeliness of *this* depending on several factors. Figure 5 shows that *this* is more likely to occur with Spanish learners (70 %) than with American journalists (50 % of the cases). The post-verbal position (corresponding to the *tagsCASE-OBLI* variable-value pair) of the proform (see Figure 6) does not impact that much as *this* is favoured in the two cases (60 % and 70 %). *This* is favoured in exophoric contexts with 80 % chances (see Figure 7). *This* is more likely to occur than *that* when anaphors are clearly identified to referents (IDTY > 65 %) or point to inferred referents (60 %). Conversely, *this* is less likely to occur (20 %) than *that* with new referents (see Figure 8).

Figure 5: The effects of the *corpus* variable values on the selection of *this* Figure 6: The effects of the *tagsCASE* variable values on the selection of *this*

Figure 7: The effects of the context variable values on the selection of *this* Figure 8: The effects of the *discourse* variable values on the selection of *this*

- 33 The model's explanatory power shows marginal $R^2=0.588$ and conditional $R^2=0.656$. Classification results indicate 0.75 accuracy compared with a 0.56 accuracy in case of computation based on proportions.

5. Discussion and Perspectives

- 34 We have shown that anaphoric processes with *it*, *this* and *that* can be analysed in terms of probabilistic trends. A mixed-effect binomial regression modelling approach helps measure the effects and the significance of linguistic factors involved in choices made by speakers. It appears that pragmatic factors are more important than morphosyntactic factors for the microsystem of pronominal anaphora. Proforms *this* and *that* both depend on the inferred nature of the referent in discourse compared with *it*, which confirms their strong potential for discourse anaphora. Exophora appears to play a significant role as speakers of both L1s prefer *this* in such contexts. These findings are in line with the literature showing that cognitive functional features such as givenness (Cornish 1999) and endophora (Fraser & Joly 1979) play an important role in selecting the forms.
- 35 The results also indicate cases in which the microsystem is learner-specific. Learners tend to favour *this* slightly more than native speakers in object positions. This confirms that learners make confusions between *this* and *that* (Lenko-Szymanska 2004 for a comparison between a native and a learner corpus). It also confirms that the two forms interact with *it* (Zhang 2015 for a comparison between the three forms within one corpus). The results of our experiment give evidence that some confusions exist between the three forms and depend on the L1. This was achieved thanks to the analysis of the three forms in two corpora.
- 36 These findings result from the comparison of two written corpora. Both corpora are of the same written mode but differ in their genres, i.e. journalistic articles and argumentative essays. There has been some debate over the influence of grammar on genres, including co-reference chains (Schnedecker 2018 for a review). By comparing the proform microsystem between two different genres, it may be argued that variations are due to this difference rather than other factors. By choosing two written corpora that include opinion texts we tried to limit this influence. However, the type of task may be more problematic since it varies between the two corpora. Writing press articles on various financial topics is not the same as writing short essays on personal experiences and the Spanish school system. Further work should focus on controlling the task variable across the corpora used for comparison (Callies, 2015).
- 37 Our approach raises the question of the type of annotation scheme to be used for anaphora. As our purpose is to analyse the microsystem formed by the three referring expressions, we are only interested in identifying the contextual features of the anaphors. More elaborate co-reference resolution schemes rely on establishing the link between anaphor and referent, e.g. GATE (Cunningham, 2002) and Glozz (Widlöcher and Mathet, 2012). To do so, the general procedure is to identify the anaphors, their relationships with co-referents and their semantic and pragmatic features (McEnery *et al*, 2006, 39). In this respect our annotation scheme is more simple as it does not include information on referring links between text units. It only focuses on the identification of the anaphor with some features. In our view, the anaphor governs its interpretation (Cornish 1999, 68) and it is its close co-text and context that include the factors of its

occurrence. Therefore, we decided to limit our annotation scheme to the properties of the referring expressions. In our framework these properties are converted into variables that are assessed with statistical methods. The results give an insight into the anaphor microsystem as they quantify the effects of the features. Corpus-based observations inform on the linguistic theorisation of the anaphoric microsystem.

- 38 As corpora increase in size, there is a need to find scalable solutions that help annotate the many dimensions of referring expressions. Automated tools appear to ensure fast and systematic tagging on the basis of methods that can be evaluated, analysed and improved. However, they show their limits in the case of co-reference annotation. Many computational methods focus mainly on co-reference resolution between tokens in a text (Lappin, 2005; Sobha *et al*, 2009; Hendrickx *et al*, 2011). They rely mainly on lexicon and morphosyntactic features to compute referential links. More recent research has focused on the automatic annotation of pragmatic features. This raises the issue of operationalising concepts that lend well to theoretical interpretation but that do not translate easily into implementable solutions. Salience is one such feature, and it has been operationalised on the basis of frequencies of references and other parameters (Landragin, 2011). This approach supports consistent and objective measurements of salience. Similarly, *givenness* has been operationalised on the basis of syntactic patterns. Syntactic features are extracted from parsed texts and exploited in a supervised learning task to assign the different states of givenness (Komen 2013). These examples show that automated solutions exist in order to add more features to the representation of anaphoric expressions. Nevertheless, performance rates advocate for manual verification. With such information, anaphoric processes will be better described, which will in turn support more comprehensive analyses of this multi-dimensional problem.
- 39 In terms of research method, the mixed-effect binomial logistic regression, as implemented in R, divides the microsystem interactions into three different models, which appears to be a limitation. By assessing binary categorical outcomes the different models overlook the weight of each missing response variable. The implementation of a multinomial logistic regression method within an R package would fill the gap. It would allow researchers to handle complex microsystems in which several forms are in paradigmatic competition.
- 40 Ultimately, the modelling approach can be exploited in two directions. By adopting a model explanation orientation, provided that goodness-of-fit indicators are satisfactory (e.g. a high R^2 indicator), we can explain and measure the effects of the factors that influence speakers. The other, complementary, orientation is to use these models as Artificial Intelligence modules implemented in Computer-Aided Language Learning (CALL) systems. The input of learners could be compared with pre-established knowledge encapsulated in models in the form of specific metrics. The system could give recommendations on the basis of predictions given by the model.

BIBLIOGRAPHY

- ARIEL M., 1988, "Referring and Accessibility", *Journal of Linguistics* 24.1, 65–87.
- BARLOW M., 2005, "Computer-Based Analyses of Learner Corpora", in R. Ellis & G. Barkhuizen (eds), *Analysing Learner Language*, Oxford, Oxford University Press, 337–57.
- BIBER D., JOHANSON S., LEECH G., CONRAD S. & FINEGAN E., 1999, *Longman Grammar of Spoken and Written English*, Harlow, Longman.
- BRESNAN J., CUENI A., NIKITA T. & BAAYERN R. H., 2007, "Predicting the Dative Alternation", in B. Gerlof, I. Kramer & J. Swarts (eds), *Cognitive Foundations of Interpretation*, Amsterdam, Royal Netherlands Academy of Arts and Sciences, 69–94.
- BURNHAM K. P. & ANDERSON D. R., 2004, "Multimodel Inference: Understanding AIC and BIC in Model Selection", *Sociological Methods & Research* 33.2, 261–304.
- CALLIES M., 2015, "Learner Corpus Methodology." In *The Cambridge Handbook of Learner Corpus Research*, in S. Granger, G. Gilquin, & F. Meunier (eds), 35–56. Cambridge: Cambridge University Press.
- CORNISH F., 1999, *Anaphora, Discourse, and Understanding. Evidence from English and French*, Oxford, Oxford University Press.
- CUNNINGHAM H., 2002, "GATE, a General Architecture for Text Engineering." *Computers and the Humanities* 36 (2): 223–54.
- DÍAZ-NEGRILLO A., 2007, "A fine-grained error tagger for learner corpora", PhD thesis, University of Jaen, Jaen, Spain.
- DÍAZ-NEGRILLO A., BALLIER N. & THOMPSON P. (eds), 2013, *Automatic Treatment and Analysis of Learner Corpus Data*, *Studies in Corpus Linguistics* 59, Amsterdam, John Benjamins.
- FRASER T. & JOLY A., 1979, « Le Système de la deixis - Esquisse d'une théorie d'expression en anglais », *Modèles Linguistiques* 1.2, 97–157.
- GAILLAT T., 2016. *Reference in Interlanguage: The Case of This and That. From Linguistic Annotation to Corpus Interoperability*, PhD Thesis in Linguistics, Université Paris-Diderot.
- GAILLAT T., 2013, "This and that in native and learner English: From typology of use to tagset characterisation." In S. Granger, G. Gilquin & F. Meunier (eds), *Corpora and Language in Use Twenty years of learner research: looking back, moving ahead*, Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain, 167–77.
- GRANGER S., 1996, "From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora", in K. Aijmer, B. Altenberg & M. Johansson (eds), *Languages in Contrast. Text-Based Cross-Linguistic Studies*, Lund, Lund University Press, 37–51.
- GRANGER S., 1994, "The learner corpus: A revolution in applied linguistics", *English Today* 10 (39.3), 25–29.
- GRANGER S., GILQUIN G. & MEUNIER F. (eds), 2015, *The Cambridge Handbook of Learner Corpus Research*, Cambridge, Cambridge University Press.
- GRIES S. T., 2015, "The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models", *Corpora* 10.1, 95–125.

- GRIES S. T., 2013, *Statistics for Linguistics with R: A Practical Introduction*, 2nd edition, Berlin, Mouton De Gruyter.
- GRIES S. T., 2008, "Corpus-based methods in analysis of second language acquisition data", in P. Robinson & N. C. Ellis (eds), *Handbook of Cognitive Linguistics and Second Language Acquisition*, New York, Routledge, 406–31.
- GRIES S. T., 2003, *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*, London, Bloomsbury.
- GUNDEL J. K., HEDBERG N. & ZACHARSKI R., 1993, "Cognitive status and the form of referring expressions in discourse", *Language* 69.2, 274–307.
- HALLIDAY M. A. K. & HASAN R., 1976, *Cohesion in English*, English Language Series, Harlow, Pearson Education.
- HENDRICKX I., SOBHA L. D., BRANCO A. & MITKOV R. (eds), 2011, *Anaphora Processing and Applications: 8th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2011*, Faro Portugal, October 6-7, 2011, *Revised Selected Papers*, Lecture Notes in Artificial Intelligence, Berlin, Springer-Verlag.
- KLEIBER G., 1992, « Anaphore-deixis : Deux approches concurrentes », in L. Danon-Boileau & M.-A. Morel (éds), *La Deixis*, Paris, Presses Universitaires de France, 613–26.
- KOMEN E. R., 2013, "Predicting Referential States Using Enriched Texts", in F. Mambrini, M. Passarotti & C. Sporleder (eds), *Proceedings of the Third Workshop on Annotation of Corpora for Research in the Humanities (ACRH-3)*, Sofia, The institute of information and communication technologies, Bulgarian academy of sciences, 49–60.
- LANDRAGIN F., 2011. « Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits. » *Corpus*, no. 10: 61–80.
- LAPAIRE J.-R. & ROTGE W., 1991, *Linguistique et grammaire de l'anglais*, Toulouse, Presses universitaires du Mirail.
- LAPPIN S., 2005, "A sequenced model of anaphora and ellipsis resolution", in A. Branco, T. McEnery & R. Mitkov (eds), *Anaphora Processing, Linguistic, Cognitive and Computational Modelling*, Current Issues in Linguistic Theory, IV, Amsterdam, John Benjamins, 3–16.
- LENKO-SZYMANSKA A., 2004, "Demonstratives as anaphora markers in advanced learners' English", in G. Aston, S. Bernardini & D. Stewart (eds), *Corpora and Language Learners*, *Studies in Corpus Linguistics* 17, Amsterdam, John Benjamins, 84–108.
- LIANG X., 2009, "A corpus-based study of developmental stages of demonstratives in Chinese English majors' writing", *Asian Social Science* 5.11, 117–125.
- MARCUS M. P., MARCINKIEWICZ M. A. & SANTORINI B., 1993, "Building a large annotated corpus of English: The Penn Treebank", *Computational Linguistics* 19.2, 313–330.
- MCENERY T., XIAO R. & TONO Y., 2006, *Corpus-Based Language Studies: An Advanced Resource Book*. Routledge Applied Linguistics. New York: Routledge.
- PETCH-TYSON S., 2000, "Demonstrative expressions in argumentative discourse", in S. Botley & A. M. McEnery (eds), *Corpus-Based and Computational Approaches to Discourse Anaphora*, *Studies in Corpus Linguistics* 3, Amsterdam, John Benjamins, 43–64.
- R CORE TEAM, 2012, *R: A Language and Environment for Statistical Computing*, Vienna, R Foundation for Statistical Computing.

- SCHNEDECKER C., 2018, "Reference Chains and Genre Identification." In *The Grammar of Genres and Styles: From Discrete to Non-Discrete Units*, in D. Legallois, T. Charnois, and M. Larjavaara (eds), 39–66. Berlin, Germany: Walter de Gruyter GmbH & Co KG.
- SCHMID H., 1994, "Probabilistic part-of-speech tagging using decision trees", in *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, 14–16.
- SCOTT K., 2013, "This and that: A procedural analysis", *Lingua* 131 (Supplement C), 49–65.
- SELINKER L., 1972, "Interlanguage", *International Review of Applied Linguistics in Language Teaching* 10.3, 209–241.
- SOBHA L. D., BRANCO A. & MITKOV R. (eds), 2009, *Anaphora Processing and Applications: 7th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2009*, Goa, India, November 5-6, 2009, *Proceedings*, Lecture Notes in Artificial Intelligence, Berlin, Springer-Verlag.
- STRAUSS S., 2002, "This, that, and it in spoken American English - a demonstrative system of gradient focus", *Language Sciences* 24.2, 131–152.
- WANG Y., MELTON G. B. & PAKHOMOV S., 2011, "It's about *this* and *that*: a description of anaphoric expressions in clinical Text", *AMIA Annual Symposium Proceedings 2011*, 1471–1480.
- WIDLÖCHER A., & MATHET Y., 2012, "The Glozz Platform: A Corpus Annotation and Mining Tool." In *Proceedings of the 2012 ACM Symposium on Document Engineering*, 171–180. DocEng '12. New York, NY, USA: ACM.
- WULFF S., ROMER U. & SWALES J., 2012, "Attended/unattended *this* in academic student writing: quantitative and qualitative perspectives", *Corpus Linguistics and Linguistic Theory* 8.1, 129–157.
- ZHANG J., 2015, "An analysis of the use of demonstratives in argumentative discourse by Chinese EFL learners", *Journal of Language Teaching and Research* 6.2, 460–465.

APPENDIXES

Appendix

The following tables show the results obtained for each model. The estimates are given in log odds with confidence Intervals (CI) and p-values for significance.

Table A.1: *It/this* mixed-effects model

Variables	est	2.5 %	97.5 %	p-values
(Intercept)	-4.766	-6.21	-3.32	9.45e-11
corpusNOCE	1.576	-3.26e-01	3.48	1.04e-01
contextEXO	21.587	-4.39e+04	4.39e+04	9.99e-01
discourseINFRD	3.733	2.20	5.27	1.89e-06***
discourseNEW	-22.022	-3.59e+05	3.59e+05	1.00
discourseONFOC	-37.255	-4.70e+04	4.69e+04	9.99e-01

tagsCASEOBLI	2.247	8.12e-01	3.68	2.15e-03**
corpusNOCE:contextEXO	-35.783	-4.49e+04	4.48e+04	9.99e-01
corpusNOCE:discourseINFRD	-0.787	-2.77e+00	1.20e+00	4.37e-01
corpusNOCE:discourseNEW	22.937	-3.59e+05	3.59e+05	1.00e+00
corpusNOCE:discourseONFOC	82.090	-1.65e+06	1.65e+06	1.00e+00
corpusNOCE:tagsCASEOBLI	-2.058	-3.85e+00	-2.70e-01	2.41e-02*

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table A.2: *It/that* mixed-effects model

Variables	est	2.5 %	97.5 %	p-values
(Intercept)	-3.930	-4.99e+00	-2.870	3.56e-13
corpusNOCE	-0.556	-1.60e+00	0.485	2.95e-01
contextEXO	3.245	-2.58e-02	6.515	5.18e-02
discourseINFRD	3.417	2.34e+00	4.489	4.19e-10***
discourseNEW	2.675	9.67e-01	4.384	2.15e-03**
discourseONFOC	-23.402	-1.03e+03	980.100	9.64e-01
tagsCASEOBLI	0.475	-4.98e-01	1.448	3.39e-01

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table A.3: *This/that* mixed-effects model

Variables	est	2.5 %	97.5 %	p-values
(Intercept)	-0.181	-0.189	-0.173	0.000
corpusNOCE	1.197	1.188	1.205	0.000
contextEXO	0.803	0.795	0.811	0.000
discourseINFRD	-0.242	-0.250	-0.234	0.000
discourseNEW	-2.251	-2.259	-2.243	0.000
discourseONFOC	14.643	-1992.360	2021.647	0.989
tagsCASEOBLI	0.373	0.365	0.381	0.000

Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

NOTES

1. Available for download at <https://github.com/tgaillat>
2. For replicability purposes, we have made the dataset available at <https://github.com/tgaillat>
3. The Akaike Information Criterion (AIC) gives an estimation of the relative quality of statistical models for a given set of data.

ABSTRACTS

This paper focuses on the use of anaphoric constructions by learners of English. The objective is to propose a comparative analysis of *it*, *this* and *that* proforms between native English speakers (NS) and Spanish speakers learning English as a foreign language. Our analysis relies on a multi-corpus approach in which texts from two corpora underwent a multi-level tagging process including morphosyntactic and pragmatic annotation with situational vs contextual and given vs new information distinctions. A dataset containing 428 occurrences of *it*, *this* and *that* was automatically built with features from the annotation layers. We conducted three pairwise binomial logistic regression modelling analyses in order to identify factors in the selection of the three forms. Results show that some pragmatic features are significant factors in the way speakers select the forms. The fact that a proform refers to a new or inferred entity in the context plays a key role in its selection. Concerning learner-specific factors, the subject position of the proform *this* influences its selection.

Cet article aborde la question des constructions anaphoriques chez les apprenants de l'anglais. L'objectif est de mener une analyse comparative des proformes *it*, *this* et *that* entre des locuteurs

natifs de langue anglaise et des apprenants d'anglais de L1 espagnol. Notre analyse se fonde sur une étude mettant en jeu deux corpus étiquetés par application de plusieurs couches d'annotation. Le schéma d'annotation comprend un étiquetage morphosyntaxique et pragmatique en intégrant les distinctions exophore/endophore et information nouvelle ou connue. Un jeu de données de 428 occurrences des trois proformes est automatiquement constitué à partir de l'extraction des formes et de leurs traits figurant dans les différentes couches d'annotation. Trois analyses par régression logistique binomiale sont effectuées sur les trois paires de proformes afin d'identifier les facteurs de sélection des formes. Les résultats des analyses montrent que les traits d'ordre pragmatique jouent un rôle significatif. Le renvoi à une entité nouvelle ou inférée est un facteur déterminant de sélection. En outre, la position sujet de *this* proforme apparaît comme un facteur spécifique aux apprenants.

INDEX

Mots-clés: anaphore, anglais L2, annotation, linguistique quantitative, comparabilité entre corpus

Keywords: anaphora, English as a Second Language (ESL), annotation, quantitative linguistics, corpus comparability

AUTHOR

THOMAS GAILLAT

Université de Rennes 1 & Insight Centre for Data Analytics NUI Galway, Irlande