Automating Second Language Acquisition Research: Integrating Information Visualisation and Machine Learning

Helen Yannakoudakis
Computer Laboratory
University of Cambridge
United Kingdom
Helen.Yannakoudakis@cl.cam.ac.uk

Ted Briscoe
Computer Laboratory
University of Cambridge
United Kingdom
Ted.Briscoe@cl.cam.ac.uk

Theodora Alexopoulou
DTAL
University of Cambridge
United Kingdom
ta259@cam.ac.uk

Abstract

We demonstrate how data-driven approaches to learner corpora can support Second Language Acquisition research when integrated with visualisation tools. We present a visual user interface supporting the investigation of a set of linguistic features discriminating between pass and fail 'English as a Second or Other Language' exam scripts. The system displays directed graphs to model interactions between features and supports exploratory search over a set of learner scripts. We illustrate how the interface can support the investigation of the co-occurrence of many individual features, and discuss how such investigations can shed light on understanding the linguistic abilities that characterise different levels of attainment and, more generally, developmental aspects of learner grammars.

1 Introduction

The Common European Framework of Reference for Languages (CEFR)[1] is an international benchmark of language attainment at different stages of learning. The English Profile (EP)[2] research programme aims to enhance the learning, teaching and assessment of English as an additional language by creating detailed reference level descriptions of the language abilities expected at each level. As part of our research within that framework, we modify and combine techniques developed for information visualisation with methodologies from computational linguistics to support a novel and more empirical perspective on CEFR levels. In particular, we build a visual user interface (hereafter UI) which aids the development of hypotheses about learner grammars using graphs of linguistic features discriminating pass/fail exam scripts for intermediate English.

Briscoe et al. (2010) use supervised discriminative machine learning methods to automate the assessment of 'English as a Second or Other Language' (ESOL) exam scripts, and in particular, the First Certificate in English (FCE) exam, which assesses English at an upper-intermediate level (CEFR level B2). They use a binary discriminative classifier to learn a linear threshold function that best discriminates passing from failing FCE scripts, and predict whether a script can be classified as such. To facilitate learning of the classification function, the data should be represented appropriately with the most relevant set of (linguistic) features. They found a discriminative feature set includes, among other feature types, lexical and part-of-speech (POS) ngrams. We extract the discriminative instances of these two feature types and focus on their linguistic analysis[3]. Table 1 presents a small subset ordered by discriminative weight.

The investigation of discriminative features can offer insights into assessment and into the linguistic properties characterising the relevant CEFR level. However, the amount and variety of data potentially made available by the classifier is considerable, as it typically finds hundreds of thousands of discriminative feature instances. Even if investigation is restricted to the most discriminative ones, calculations of relationships be-

[1] http://www.coe.int/t/dg4/linguistic/cadre_en.asp
[2] http://www.englishprofile.org/

[3] Briscoe et al. (2010) POS tagged and parsed the data using the RASP toolkit (Briscoe et al., 2006). POS tags are based on the CLAWS tagset.

Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, pages 35–43,
Avignon, France, April 23 - 24 2012.

tween features can rapidly grow and become overwhelming. Discriminative features typically capture relatively low-level, specific and local properties of texts, so features need to be linked to the scripts they appear in to allow investigation of the contexts in which they occur. The scripts, in turn, need to be searched for further linguistic properties in order to formulate and evaluate higher-level, more general and comprehensible hypotheses which can inform reference level descriptions and understanding of learner grammars.

The appeal of information visualisation is to gain a deeper understanding of important phenomena that are represented in a database (Card et al., 1999) by making it possible to navigate large amounts of data for formulating and testing hypotheses faster, intuitively, and with relative ease. An important challenge is to identify and assess the usefulness of the enormous number of projections that can potentially be visualised. Exploration of (large) databases can lead quickly to numerous possible research directions; lack of good tools often slows down the process of identifying the most productive paths to pursue.

In our context, we require a tool that visualises features flexibly, supports interactive investigation of scripts instantiating them, and allows statistics about scripts, such as the co-occurrence of features or presence of other linguistic properties, to be derived quickly. One of the advantages of using visualisation techniques over command-line database search tools is that Second Language Acquisition (SLA) researchers and related users, such as assessors and teachers, can access scripts, associated features and annotation intuitively without the need to learn query language syntax.

We modify previously-developed visualisation techniques (Di Battista et al., 1999) and build a visual UI supporting hypothesis formation about learner grammars. Features are grouped in terms of their co-occurrence in the corpus and directed graphs are used in order to illustrate their relationships. Selection of different feature combinations automatically generates queries over the data and returns the relevant scripts as well as associations with meta-data and different types of errors committed by the learners[4]. In the next sections we describe in detail the visualiser, illustrate how it can support the investigation of individual features, and discuss how such investigations can shed light on the relationships between features and developmental aspects of learner grammars.

To the best of our knowledge, this is the first attempt to visually analyse as well as perform a linguistic interpretation of discriminative features that characterise learner English. We also apply our visualiser to a set of 1,244 publically-available FCE ESOL texts (Yannakoudakis et al., 2011) and make it available as a web service to other researchers[5].

Feature	Example
VM_RR (POS bigram: $+$)	*could clearly*
,_because (word bigram: $-$)	*, because of*
necessary (word unigram: $+$)	*it is necessary that*
the_people (word bigram: $-$)	**the people are clever*
VV∅_VV∅ (POS bigram: $-$)	**we go see film*
NN2_VVG (POS bigram: $+$)	*children smiling*

Table 1: Subset of features ordered by discriminative weight; $+$ and $-$ show their association with either passing or failing scripts.

2 Dataset

We use texts produced by candidates taking the FCE exam, which assesses English at an upper-intermediate level. The FCE texts, which are part of the Cambridge Learner Corpus[6], are produced by English language learners from around the world sitting Cambridge Assessment's ESOL examinations[7]. The texts are manually tagged with information about linguistic errors (Nicholls, 2003) and linked to meta-data about the learners (e.g., age and native language) and the exam (e.g., grade).

3 The English Profile visualiser

3.1 Basic structure and front-end

The English Profile (EP) visualiser is developed in Java and uses the Prefuse library (Heer et al., 2005) for the visual components. Figure 1 shows its front-end. Features are represented

[4]Our interface integrates a command-line Lucene search tool (Gospodnetic and Hatcher, 2004) developed by Gram and Buttery (2009).

[5]Available by request: http://ilexir.co.uk/applications/ep-visualiser/

[6]http://www.cup.cam.ac.uk/gb/elt/catalogue/subject/custom/item3646603/

[7]http://www.cambridgeesol.org/

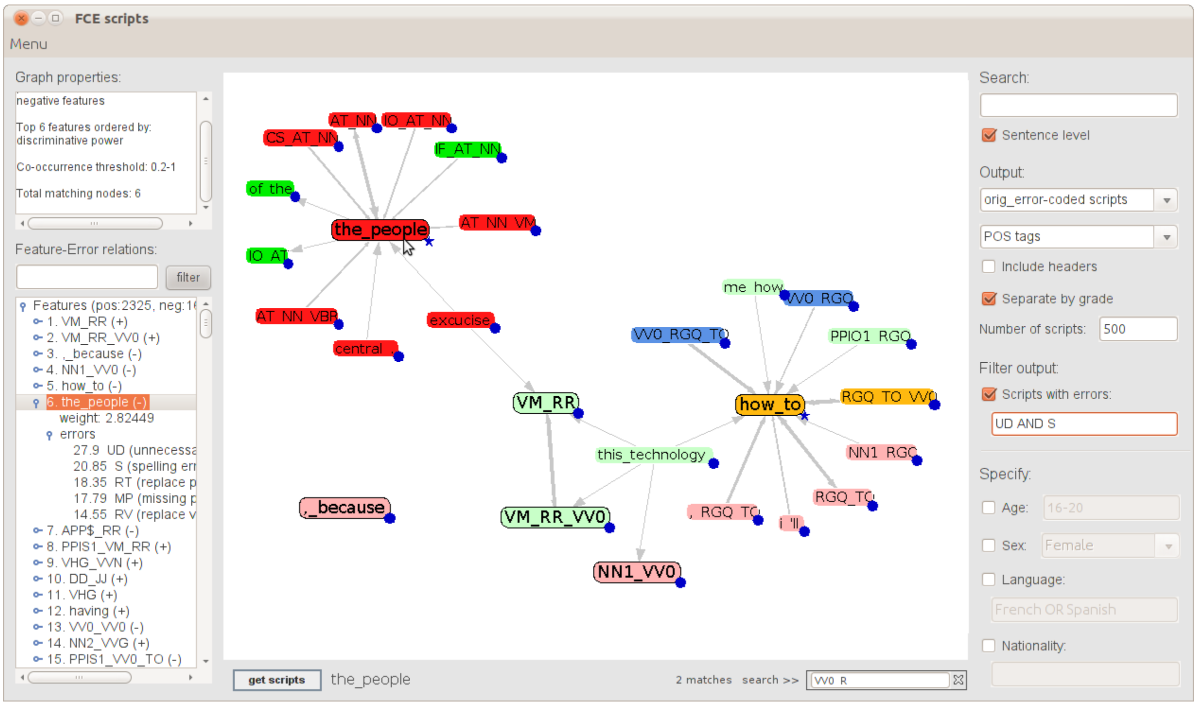


Figure 1: Front-end of the EP visualiser.

by a labelled node and displayed in the central panel; positive features (i.e., those associated with passing the exam) are shaded in a light green colour while negative ones are light red[8]. A field at the bottom right supports searching for features/nodes that start with specified characters and highlighting them in blue. An important aspect is the display of feature patterns, discussed in more detail in the next section (3.2).

3.2 Feature relations

Crucial to understanding discriminative features is finding the relationships that hold between them. We calculate co-occurrences of features at the sentence-level in order to extract 'meaningful' relations and possible patterns of use. Combinations of features that may be 'useful' are kept while the rest are discarded. 'Usefulness' is measured as follows:

Consider the set of all the sentences in the corpus $S = \{s_1, s_2, ..., s_N\}$ and the set of all the features $F = \{f_1, f_2, ..., f_M\}$. A feature $f_i \in F$ is associated with a feature $f_j \in F$, where $i \neq j$ and $1 \leq i, j \leq M$, if their relative co-occurrence score is within a predefined range:

$$\text{score}(f_j, f_i) = \frac{\sum_{k=1}^{N} \text{exists}(f_j, f_i, s_k)}{\sum_{k=1}^{N} \text{exists}(f_i, s_k)} \quad (1)$$

[8] Colours can be customised by the user.

where $s_k \in S$, $1 \leq k \leq N$, exists() is a binary function that returns 1 if the input features occur in s_k, and $0 \leq \text{score}(f_j, f_i) \leq 1$. We group features in terms of their relative co-occurrence within sentences in the corpus and display these co-occurrence relationships as directed graphs. Two nodes (features) are connected by an edge if their score, based on Equation (1), is within a user-defined range (see example below). Given f_i and f_j, the outgoing edges of f_i are modelled using $\text{score}(f_j, f_i)$ and the incoming edges using $\text{score}(f_i, f_j)$. Feature relations are shown via highlighting of features when the user hovers the cursor over them, while the strength of the relations is visually encoded in the edge width.

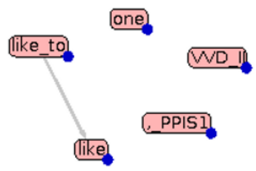
For example, one of the highest-weighted positive discriminative features is VM_RR (see Table 1), which captures sequences of a modal auxiliary followed by an adverb as in *will always (avoid)* or *could clearly (see)*. Investigating its relative co-occurrence with other features using a score range of 0.8–1 and regardless of directionality, we find that VM_RR is related to the following: (i) POS ngrams: RR_VB∅_AT1, VM_RR_VB∅, VM_RR_VH∅, PPH1_VM_RR, VM_RR_VV∅, PPIS1_VM_RR, PPIS2_VM_RR, RR_VB∅; (ii) word ngrams: will_also, can_only, can_also, can_just. These relations show us the syntactic environments of the feature (i) or its characteristic lexicalisations (ii).

3.3 Dynamic creation of graphs via selection criteria

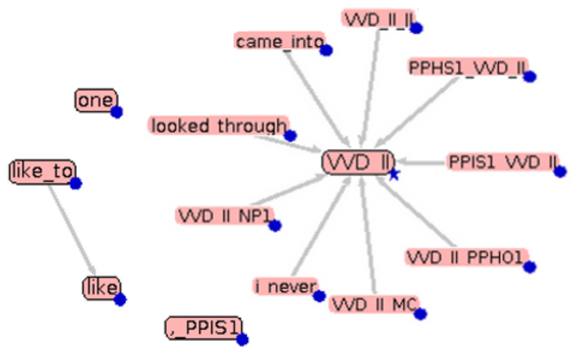
Questions relating to a graph display may include information about the most connected nodes, separate components of the graph, types of interconnected features, etc. However, the functionality, usability and tractability of graphs is severely limited when the number of nodes and edges grows by more than a few dozen (Fry, 2007). In order to provide adequate information, but at the same time avoid overly complex graphs, we support dynamic creation and visualisation of graphs using a variety of selection criteria. The EP visualiser supports the flexible investigation of the top 4,000 discriminative features and their relations.

The *Menu* item on the top left of the UI in Figure 1 activates a panel that enables users to select the top N features to be displayed. The user can choose whether to display positive and/or negative features and set thresholds for, as well as rank by discriminative weight, connectivity with other features (i.e., the number of features it is connected to), and frequency. For instance, a user can choose to investigate features that have a connectivity between 500 and 900, rank them by frequency and display the top 100. Highly-connected features might tell us something about the learner grammar while infrequent features, although discriminative, might not lead to useful linguistic insights. Additionally, users can investigate feature relations and set different score ranges according to Equation (1), which controls the edges to be displayed.

Figure 2(a) presents the graph of the 5 most frequent negative features, using a score range of 0.8–1. The system displays only one edge, while the rest of the features are isolated. However, these features might be related to other features from the list of 4,000 (which are not displayed since they are not found in the top N list of features). Blue aggregation markers in the shape of a circle, located at the bottom right of each node, are used to visually display that information. When a node with an aggregation marker is selected, the system automatically expands the graph and displays the related features. The marker shape of an expanded node changes to a star, while a different border stroke pattern



(a) Graph of the top 5 most frequent negative features using a score range of 0.8–1.



(b) Expanded graph when the aggregation marker for the feature VVD_II is selected.

Figure 2: Dynamic graph creation.

is used to visually distinguish the revealed nodes from the top N. Figure 2(b) presents the expanded graph when the aggregation marker for the feature VVD_II is selected. If the same aggregation marker is selected twice, the graph collapses and returns to its original form.

3.4 Feature–Error relations

The FCE texts have been manually error-coded (Nicholls, 2003) so it is possible to find associations between discriminative features and specific error types. The *Feature–Error relations* component on the left of Figure 1 displays a list of the features, ranked by their discriminative weight, together with statistics on their relations with errors. Feature–error relations are computed at the sentence level by calculating the proportion of sentences containing a feature that also contain a specific error (similar to Equation (1)). In the example in Figure 1, we see that 27% of the sentences that contain the feature bigram the_people also have an unnecessary determiner (UD) error, while 14% have a replace verb (RV) error[9].

[9]In the example image we only output the top 5 errors (can be customised by the user).

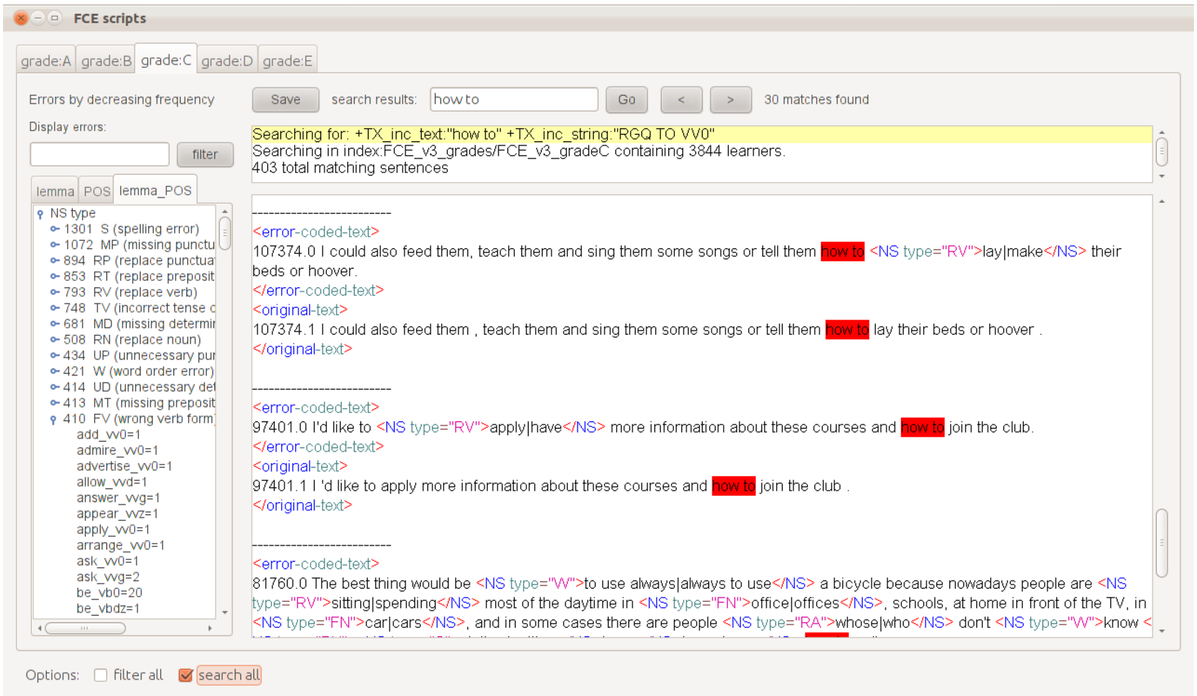


Figure 3: Sentences, split by grade, containing occurrences of how_to and RGQ_TO_VV∅. The list on the left gives error frequencies for the matching scripts, including the frequencies of lemmata and POSs inside an error.

3.5 Searching the data

In order to allow the user to explore how features are related to the data, the EP visualiser supports browsing operations. Selecting multiple features – highlighted in yellow – and clicking on the button *get scripts* returns relevant scripts. The right panel of the front-end in Figure 1 displays a number of search and output options. Users can choose to output the original/error-coded/POS-tagged text and/or the grammatical relations found by the RASP parser (Briscoe et al., 2006), while different colours are used in order to help readability. Data can be retrieved at the sentence or script level and separated according to grade. Additionally, Boolean queries can be executed in order to examine occurrences of (selected features and) specific errors only[10]. Also, users can investigate scripts based on meta-data information such as learner age.

Figure 3 shows the display of the system when the features how_to and RGQ_TO_VV∅ (*how to* followed by a verb in base form) are selected. The text area in the centre displays sentences instantiating them. A search box at the top supports navigation, highlighting search terms in red, while a small text area underneath displays the current search query, the size of the database and the number of matching scripts or sentences. The *Errors by decreasing frequency* pane on the left shows a list of the errors found in the matching scripts, ordered by decreasing frequency. Three different tabs (lemma, POS and lemma_POS) provide information about and allow extraction of counts of lemmata and POSs inside an error tag.

3.6 Learner native language

Research on SLA highlights the possible effect of a native language (L1) on the learning process. Using the *Menu* item on the top left corner of Figure 1, users can select the language of interest while the system displays a new window with an identical front-end and functionality. Feature–error statistics are now displayed per L1, while selecting multiple features returns scripts written by learners speaking the chosen L1.

4 Interpreting discriminative features: a case study

We now illustrate in greater depth how the EP visualiser can support interpretation of discriminative features: the POS trigram RG_JJ_NN1 (–) is

[10]For example, users can activate the *Scripts with errors:* option and type 'R OR W'. This will return sentences containing replace or word order errors.

the 18th most discriminative (negative) feature. It corresponds to a sequence of a degree adverb followed by an adjective and a singular noun as in *very good boy*. The question is why such a feature is negative since the string is not ungrammatical. Visualisation of this feature using the 'dynamic graph creation' component of the visualiser allows us to see the features it is related to. This offers an intuitive and manageable way of investigating the large number of underlying discriminative features.

We find that RG_JJ_NN1 is related to its discriminative lexicalisation, very_good (−), which is the 513th most discriminative feature. Also, it is related to JJ_NN1_II (−) (e.g., *difficult sport at*), ranked 2,700th, which suggests a particular context for RG_JJ_NN1 when the noun is followed by a preposition. Searching for this conjunction of features in scripts, we get production examples like *1a,b,c*. Perhaps more interestingly, RG_JJ_NN1 is related to VBZ_RG (−) (ranked 243rd): *is* followed by a degree adverb. This relation suggests a link with predicative structures since putting the two ngrams together yields strings VBZ_ RG_JJ_NN1 corresponding to examples like *1c,d*; if we also add _II we get examples like *1c*.

1a *It might seem to be* ***very difficult sport at*** *the beginning.*

1b *We know a lot about* ***very difficult situation in*** *your country.*

1c *I think it'****s very good idea to*** *spending vacation together.*

1d *Unix* ***is very powerful system*** *but there is one thing against it.*

The associations between features already give an idea of the source of the problem. In the sequences including the verb *be* the indefinite article is omitted. So the next thing to investigate is if indeed RG_JJ_NN1 is associated with article omission, not only in predicative contexts, but more generally. The *Feature–Error relations* component of the UI reveals an association with MD (missing determiner) errors: 23% of sentences that contain RG_JJ_NN1 also have a MD error. The same holds for very_good, JJ_NN1_II and VBZ_RG with percentages 12%, 14% and 15% respectively. We then compared the number of MD errors per script across different types of scripts. Across all scripts the ratio MD:doc is 2.18, that is, approximately 2 MD errors per script; in RG_JJ_NN1 scripts this ratio goes up to 2.75, so that each script has roughly 3 MD errors. VBZ_RG follows with 2.68, JJ_NN1_II with 2.48, and very_good with 2.32. In scripts containing all features the ratio goes up to 4.02 (3.68 without very_good), and in scripts containing VBZ_RG_JJ the ratio goes up to 2.73. Also, in most of these scripts the error involves the indefinite article. The emerging picture then is that there is a link between these richer nominal structures that include more than one modifier and the omission of the article. Two questions arise: (i) why these richer nominals should associate with article omission and (ii) why only singular nouns are implicated in this feature.

Language	f_1	f_2	f_3	f_4
all	0.26	0.40	0.02	0.03
Turkish	0.29	0.48	0.04	0.03
Japanese	0.17	0.39	0.02	0.02
Korean	0.30	0.58	0.06	0.03
Russian	0.35	0.52	0.03	0.03
Chinese	0.25	0.56	0.02	0.03
French	0.21	0.41	0.00	0.03
German	0.19	0.41	0.00	0.02
Spanish	0.27	0.32	0.00	0.03
Greek	0.30	0.35	0.02	0.02

Table 2: $f_{1/2/3/4}$:doc ratios for different L1s.

Article omission errors are typical of learners coming from L1s lacking an article system (Robertson, 2000; Ionin and Montrul, 2010; Hawkins and Buttery, 2010). Trenkic (2008) proposes that such learners analyse articles as adjectival modifiers rather than as a separate category of determiners or articles. When no adjective is involved, learners may be aware that bare nominals are ungrammatical in English and provide the article. However, with complex adjectival phrases, learners may omit the article because of the presence of a degree adverb. In order to evaluate this hypothesis further we need to investigate if article omission is indeed more pronounced in our data with more complex adjectival phrases e.g., *very difficult situation* than with simpler ones e.g., *nice boy* and whether this is primarily the case for learners from L1s lacking articles.

Again, using the *Errors by decreasing frequency* pane we found that the MD:doc ratio in scripts containing the bigram JJ_NN1 is 2.20. Additionally, in scripts containing JJ_NN1 and not RG_JJ_NN1 it goes down to 2.04. These results are much lower compared to the MD:doc ratio in scripts containing RG_JJ_NN1 and/or the features with which it is related (see above), further supporting our hypothesis. We also found the ratio of RG_JJ_NN1 (f_1) occurrences per document across different L1s, as well as the ratio of VBZ_RG_JJ (f_2), VBZ_RG_JJ_NN1 (f_3) and RG_JJ_NN1_II (f_4). As shown in Table 2 there is no correlation between these features and the L1, with the exception of f_1 and f_2 which are more pronounced in Korean and Russian speakers, and of f_3 which seems completely absent from French, German and Spanish which all have articles. The exception is Greek which has articles but uses bare nominals in predicative structures.

However, a more systematic pattern is revealed when relations with MD errors are considered (using the *Feature–Error relations* and *Errors by decreasing frequency* components for different L1s). As shown in Table 3, there is a sharp contrast between L1s with articles (French, German, Spanish and Greek) and those without (Turkish, Japanese, Korean, Russian, Chinese), which further supports our hypothesis. A further question is why only the singular article is implicated in this feature. The association with predicative contexts may provide a clue. Such contexts select nominals which require the indefinite article only in the singular case; compare *Unix is (a) very powerful system* with *Macs are very elegant machines*.

In sum, navigating the UI, we formed some initial interpretations for why a particular feature is negatively discriminative. In particular, nominals with complex adjectival phrases appear particularly susceptible to article omission errors by learners of English with L1s lacking articles. The example illustrates not just the usefulness of visualisation techniques for navigating and interpreting large amounts of data, but, more generally the relevance of features weighted by discriminative classifiers. Despite being superficial in their structure, POS ngrams can pick up syntactic environments linked to particular phenomena. In this case, the features do not just identify a high rate of article omission errors, but, importantly, a particular syntactic environment triggering higher rates of such errors.

	sentences%		**MD:doc**	
Language	f_1	f_2	f_1	f_2
all	23.0	15.6	2.75	2.73
Turkish	45.2	29.0	5.81	5.82
Japanese	44.4	22.3	4.48	3.98
Korean	46.7	35.0	5.48	5.31
Russian	46.7	23.4	5.42	4.59
Chinese	23.4	13.5	3.58	3.25
French	6.9	6.7	1.32	1.49
German	2.1	3.0	0.91	0.92
Spanish	10.0	9.6	1.18	1.35
Greek	15.5	12.9	1.60	1.70

Table 3: $f_{1/2}$ relations with MD errors for different L1s, where sentences% shows the proportion of sentences containing $f_{1/2}$ that also contain a MD.

5 Previous work

To the best of our knowledge, this is the first attempt to visually analyse as well as perform a linguistic interpretation of discriminative features that characterise learner English.

Collins (2010) in his dissertation addresses visualisation for NLP research. The *Bubble Sets* visualisation draws secondary set relations around arbitrary collections of items, such as a linguistic parse tree. *VisLink* provides a general platform within which multiple visualisations of language (e.g., a force-directed graph and a radial graph) can be connected, cross-queried and compared. Moreover, he explores the space of content analysis. *DocuBurst* is an interactive visualisation of document content, which spatially organizes words using an expert-created ontology (e.g., WordNet). *Parallel Tag Clouds* combine keyword extraction and coordinated visualisations to provide comparative overviews across subsets of a faceted text corpus. Recently, Rohrdantz et al. (2011) proposed a new approach to detecting and investigating changes in word senses by visually modelling and plotting aggregated views about the diachronic development in word contexts.

Visualisation techniques have been successfully used in other areas including the humanities (e.g., Plaisant et al. (2006) and Don et al. (2007)), as well as genomics (e.g., Meyer et al. (2010a) and Meyer et al. (2010b)). For example, Meyer et al. (2010a) present a system that supports the inspection and curation of data sets showing gene expression over time, in conjunction with the spatial location of the cells where the genes are expressed.

Graph layouts have been effectively used in the analysis of domains such as social networks (e.g., terrorism network) to allow for a systematic exploration of a variety of Social Network Analysis measures (e.g., Gao et al. (2009) and Perer and Shneiderman (2006)). Heer and Boyd (2005) have implemented *Vizster*, a visualisation system for the exploration of on-line social networks (e.g., facebook) designed to facilitate the discovery of people, promote awareness of community structure etc. Van Ham et al. (2009) introduce *Phrase Net*, a system that analyses unstructured text by taking as input a predefined pattern and displaying a graph whose nodes are words and whose edges link the words that are found as matches.

We believe our integration of highly-weighted discriminative features identified by a supervised classifier into a graph-based visualiser to support linguistic SLA research is, however, novel.

6 Conclusions

We have demonstrated how a data-driven approach to learner corpora can support SLA research when guided by discriminative features and augmented with visualisation tools. We described a visual UI which supports exploratory search over a corpus of learner texts using directed graphs of features, and presented a case study of how the system allows SLA researchers to investigate the data and form hypotheses about intermediate level learners. Although the usefulness of the EP visualiser should be confirmed through more rigorous evaluation techniques, such as longitudinal case studies (Shneiderman and Plaisant, 2006; Munzner, 2009) with a broad field of experts, these initial explorations are encouraging. One of the main advantages of using visualisation techniques over command-line database search tools is that SLA researchers can start developing and testing hypotheses without the need to learn a query syntax first.

We would also like to point out that we adopted a user-driven development of the visualiser based on the needs of the third author, an SLA researcher who acted as a design partner during the development of the tool and was eager to use and test it. There were dozens of meetings over a period of seven months, and the feedback on early interfaces was incorporated in the version described here. After the prototype reached a satisfactory level of stability, the final version overall felt enjoyable and inviting, as well as allowed her to form hypotheses and draw on different types of evidence in order to substantiate it (Alexopoulou et al., 2012). Future work will include the development, testing and evaluation of the UI with a wider range of users, as well as be directed towards investigation and evaluation of different visualisation techniques of machine learned or extracted features that support hypothesis formation about learner grammars.

Acknowledgments

We are grateful to Cambridge ESOL for supporting this research. We would like to thank Marek Rei, Øistein Andersen, Paula Buttery and Angeliki Salamoura for fruitful discussions and feedback, Tim Parish for making the tool available on the web, as well as the anonymous reviewers for their valuable comments and suggestions.

References

Theodora Alexopoulou, Helen Yannakoudakis, and Angeliki Salamoura. 2012. Classifying intermediate Learner English: a data-driven approach to learner corpora. *to appear*.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the COLING/ACL*, volume 6.

Ted Briscoe, Ben Medlock, and Øistein Andersen. 2010. *Automated Assessment of ESOL Free Text Examinations*. University of Cambridge, Computer Laboratory, TR-790.

Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann.

Christopher M. Collins. 2010. *Interactive Visualizations of natural language*. Ph.D. thesis, University of Toronto.

Giuseppe Di Battista, Peter Eades, Roberto Tamassia, and Ioannis G. Tollis. 1999. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall Press.

Anthony Don, Elena Zheleva, Machon Gregory, Sureyya Tarkan, Loretta Auvil, Tanya Clement, Ben Shneiderman, and Catherine Plaisant. 2007. Discovering interesting usage patterns in text collections: integrating text mining with visualization. In

Proceedings of the sixteenth ACM conference on information and knowledge management, pages 213–222. ACM.

Ben Fry. 2007. *Visualizing Data: Exploring and Explaining Data with the Processing Environment*. O'Reilly Media.

Jie Gao, Kazuo Misue, and Jiro Tanaka. 2009. A Multiple-Aspects Visualization Tool for Exploring Social Networks. *Human Interface and the Management of Information*, pages 277–286.

Otis Gospodnetic and Erik Hatcher. 2004. *Lucene in Action*. Manning Publications.

Lu Gram and Paula Buttery. 2009. A tutorial introduction to iLexIR Search. unpublished.

John Hawkins and Paula Buttery. 2010. Criterial features in Learner Corpora: theory and illustrations. *English Profile Journal*, 1(1):1–23.

Jeffrey Heer and Danah Boyd. 2005. Vizster: visualizing online social networks. *IEEE Symposium on Information Visualization (INFOVIS)*, pages 32–39.

Jeffrey Heer, Stuart K. Card, and James A. Landay. 2005. Prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 421–430, New York, USA. ACM.

Tania Ionin and Silvina Montrul. 2010. The role of l1 transfer in the interpretation of articles with definite plurals in l2 english. *Language Learning*, 60(4):877–925.

Miriah Meyer, Tamara Munzner, Angela DePace, and Hanspeter Pfister. 2010a. MulteeSum: a tool for comparative spatial and temporal gene expression data. *IEEE transactions on visualization and computer graphics*, 16(6):908–17.

Miriah Meyer, Bang Wong, Mark Styczynski, Tamara Munzner, and Hanspeter Pfister. 2010b. Pathline: A tool for comparative functional genomics. *Computer Graphics*, 29(3).

Tamara Munzner. 2009. A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6).

Diane Nicholls. 2003. The Cambridge Learner Corpus-error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference*, pages 572–581.

Adam Perer and Ben Shneiderman. 2006. Balancing Systematic and Flexible Exploration of Social Networks. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):693–700.

Catherine Plaisant, James Rose, Bei Yu, Loretta Auvil, Matthew G. Kirschenbaum, Martha N. Smith, Tanya Clement, and Greg Lord. 2006. Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 141–150. ACM.

Daniel Robertson. 2000. Variability in the use of the English article system by Chinese learners of English. *Second Language Research*, 2:135–172.

Christian Rohrdantz, Annette Hautli, Thomas Mayer, and Miriam Butt. 2011. Towards tracking semantic change by visual analytics. *Proceedings of the 49th Meeting of the Association for Computational Linguistics*, pages 305–310.

Ben Shneiderman and Catherine Plaisant. 2006. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*. ACM.

Danijela Trenkic. 2008. The representation of English articles in second language grammars: Determiners or adjectives? *Bilingualism: Language and Cognition*, 11(01):1–18.

Frank Van Ham, Martin Wattenberg, and Fernanda B. Viégas. 2009. Mapping text with phrase nets. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1169–76.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.