The English Grammar Profile

Paula Buttery
The Automated Language Teaching & Assessment (ALTA) Institute
Department of Theoretical and Applied Linguistics
University of Cambridge
The major objective of the English Profile Programme is to analyse language produced by learners of English in order to [discover] what they can and can’t do at each of the Common European Framework of Reference levels.

Nick Saville and Mike Milanovic, Cambridge English Language Assessment, 2012
Projects within the programme of work describe the gradual mastery of English across the six CEFR levels (across many L1s).

The projects are evidence based, largely informed by the Cambridge Learner Corpus.

Work in collaboration with the ALTA institute, Cambridge University Press; and Cambridge English Language Assessment.
The English Profile Programme is evidence based

### The Cambridge Learner Corpus
- 200,000 exam scripts
- 147 language backgrounds
- 217 countries.

### Cambridge English Corpus
(formerly the Cambridge International Corpus)
- multi-million words of learner English (superset of the Cambridge Learner Corpus)
- Written and Spoken language
- newspapers, the web, books, magazines, radio, exams, schools, universities, the workplace and everyday conversation.

No dedicated textbook corpus...CAE
Some general details:

- Compiled by Annette Capel
- in the region of 20,000 searchable words and phrases (across multiple levels)
- Freely available for teachers and researchers
- Text inspector tool for teachers and editors
believe /briːlv/ 

Word family:
Nouns: belief
Verbs: believe
Adjectives: unbelievable

VERB [T]

TRUE
A2 to think that something is true, or that what someone says is true

Dictionary examples:
They believe that their health has suffered because of the chemicals.
I can’t believe that she wants to go out with me.
I told you she was there but you wouldn’t believe me.

Learner example:
I believe him.

THINK
A2 to think something, without being completely sure

Dictionary examples:
"Is she coming alone?" "I believe so."
All the crew are missing, believed dead.

Learner example:
I believe it is in your bedroom.

not believe your eyes/ears
B1 to be very surprised when you see someone or something, or when you hear what someone says

Dictionary example:
I couldn’t believe my ears when Dan said they were getting married.
English Vocabulary Profile: How much vocabulary?

Search results for A1-B2 (310 matches)

Core results:
- **absolutely** adverb COMPLETELY B1
- **accidentally** adverb B2
- **accurately** adverb B1
- **actively** adverb B2
- **actually** adverb TRUTH A2
- **actually** adverb SURPRISE B1
- **actually** adverb OPPOSITE B2
- **additionally** adverb B2
- **adequately** adverb B2
- **admittedly** adverb B2
- **aggressively** adverb B2
- **alternatively** adverb B2
- **amazingly** adverb B1
- **angrily** adverb B1
- **annually** adverb B2
- **anxiously** adverb B2
- **apparently** adverb AS HEARD/READ B2
- **apparently** adverb SEEMS TRUE B2
- **appropriately** adverb B2
- **approximately** adverb B1
Teachers and students can use this page to check the difficulty of any text, by showing its content in terms of the CEFR levels of words as described in the **English Vocabulary Profile** (EVP), a free online resource.

The EVP classifies words by level according to their meaning, as some senses of a word may be learned at a higher level than others. Text inspector chooses the lowest level sense by default, but the user can click on each word in the results display and select a different sense if necessary.

Texts are limited to **500** words per document. To process larger documents and to save your data, please use the advanced site at [www.textinspector.com](http://www.textinspector.com) (subscription required).

For Help or Comments please go to: [www.textinspector.com/help](http://www.textinspector.com/help)

Copy and paste, or type, text into the box below to analyse.

Alice was not a bit hurt, and she jumped up in a moment. She looked up, but it was all dark overhead; before her was another long passage and the White Rabbit was still in sight, hurrying down it. There
English Vocabulary Profile: Text analysis

Input

alice was a not a a bit hurt and she jumped up in a a moment she looked up but it was all dark overhead before her was another long passage and the white rabbit was still in sight hurrying down it there was not a a moment to be lost away went alice like the wind and was just in time to hear it say as it turned a corner oh my ears and whiskers how late it's getting she was close behind it when she turned the corner but the rabbit was no longer to be seen

*Text Inspector* shows the lowest value of each item by default. For more accuracy, click on an item and choose the correct use from the list. Then click on the UPDATE button below to update your statistics.

Update Back
English Vocabulary Profile: Text statistics

<table>
<thead>
<tr>
<th>Word List</th>
<th>Types</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>39</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>(66.10%)</td>
<td>(73.40%)</td>
</tr>
<tr>
<td>A2</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>(22.03%)</td>
<td>(18.09%)</td>
</tr>
<tr>
<td>B1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(3.39%)</td>
<td>(2.13%)</td>
</tr>
<tr>
<td>B2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(1.69%)</td>
<td>(1.06%)</td>
</tr>
<tr>
<td>Unlisted</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>(6.78%)</td>
<td>(5.32%)</td>
</tr>
</tbody>
</table>

Statistics and details (Toggle all words)

On the main site you can process 10,000 words, and export your data. Go to www.textinspector.com. (Subscription required).
The goal of EGP is to capture the grammar that learners are using **appropriately** and **correctly** at each CEFR level.

Mike McCarthy

Some general details:
- Compiled by Anne O’Keefe and Geraldine Mark
- 1239 searchable entries
- Freely available for teachers and researchers
<table>
<thead>
<tr>
<th>SuperCategory</th>
<th>SubCategory</th>
<th>Level</th>
<th>Can-do statement</th>
<th>Example</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJECTIVES</td>
<td>combining</td>
<td>A1</td>
<td><strong>FORM:</strong> COMBINING TWO ADJECTIVES WITH 'AND'</td>
<td>Example</td>
<td>Details</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Can use 'and' to join a limited range of common adjectives.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADJECTIVES</td>
<td>modifying</td>
<td>A1</td>
<td><strong>FORM:</strong> WITH 'VERY'</td>
<td>Example</td>
<td>Details</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Can use 'very' with a limited range of common gradable adjectives.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADJECTIVES</td>
<td>position</td>
<td>A1</td>
<td><strong>FORM:</strong> ATTRIBUTIVE (WITH NOUNS)</td>
<td>Example</td>
<td>Details</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Can use a limited range of adjectives attributively, before a noun.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADJECTIVES</td>
<td>position</td>
<td>A1</td>
<td><strong>FORM:</strong> PREDICATIVE, WITH 'BE'</td>
<td>Example</td>
<td>Details</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Can use a limited range of adjectives predicatively, after 'be'.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADJECTIVES</td>
<td>superlatives</td>
<td>A1</td>
<td><strong>FORM:</strong> 'MY BEST FRIEND'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SuperCategory</td>
<td>SubCategory</td>
<td>Level</td>
<td>Can-do statement</td>
<td>Example</td>
<td>Details</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>-------</td>
<td>----------------------------------------------------------------------------------</td>
<td>---------</td>
<td>---------</td>
</tr>
</tbody>
</table>
| ADJECTIVES    | combining    | A1    | **FORM:** COMBINING TWO ADJECTIVES WITH 'AND'  
Can use 'and' to join a limited range of common adjectives. |         |         |
| ADJECTIVES    | modifying    | A1    | **FORM:** WITH 'VERY'  
Can use 'very' with a limited range of common gradable adjectives. |         |         |
| ADJECTIVES    | position     | A1    | **FORM:** ATTRIBUTIVE (WITH NOUNS)  
Can use a limited range of adjectives attributively, before a noun.  
noun phrases |         |         |
| ADJECTIVES    | position     |       | **FORM:** 'MY BEST FRIEND'  
She's my best friend. (A1 BREAKTHROUGH; 2006; Polish; Pass) |         |         |
| ADJECTIVES    | superlatives |       | Shahin is my best friend. (A1 BREAKTHROUGH; 2006; Urdu; Pass)                   |         |         |
| ADVERBS       | adverbs as modifiers |       | Can use 'soon' in the phrases 'See you soon' and 'Get well soon', as a signing-off device. |         |         |
## English Grammar Profile: Detailed entries

<table>
<thead>
<tr>
<th>Element</th>
<th>FORM:</th>
<th>PREDICATIVE, WITH 'BE'</th>
</tr>
</thead>
<tbody>
<tr>
<td>SuperCat</td>
<td>ADJECTIVES</td>
<td></td>
</tr>
<tr>
<td>SubCat</td>
<td>position</td>
<td></td>
</tr>
<tr>
<td>Lexical Range</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level</td>
<td>A1</td>
<td></td>
</tr>
<tr>
<td>Cando</td>
<td>Can use a limited range of adjectives predicatively, after 'be'.</td>
<td></td>
</tr>
<tr>
<td>Corrected Learner Example</td>
<td>I like my home because my bedroom is big. (Turkish, A1 BREAKTHROUGH, Pass, 2009)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>This place is lovely. (Polish, A1 BREAKTHROUGH, Pass, 2007)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>My kitchen is nice. (Farsi, A1 BREAKTHROUGH, Pass, 2009)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>This place is beautiful. (Tagalog, A1 BREAKTHROUGH, Pass, 2007)</td>
<td></td>
</tr>
<tr>
<td>Uncorrected Learner Example</td>
<td>I like my home because my bedroom is big and I have double bed and I have lot of pictures on the wall. In this house is near the town center. (A1 BREAKTHROUGH; 2009; Turkish; Pass)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>This place is lovely. (A1 BREAKTHROUGH; 2007; Polish; Pass)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>My kitchen is nice and it has got now cooker. (A1 BREAKTHROUGH; 2009; Farsi; Pass)</td>
<td></td>
</tr>
</tbody>
</table>
## English Grammar Profile: What grammar?

**SuperCategory**
- ADJECTIVES
- ADVERBS
- CLAUSES
- CONJUNCTIONS
- DETERMINERS
- DISCOURSE MARKERS
- FOCUS
- FUTURE
- MODALITY
- NEGATION
- NOUNS
- PASSIVES
- PAST
- PREPOSITIONS
- PRESENT
- PRONOUNS
- QUESTIONS
- REPORTED SPEECH
- VERBS

**SubCategory**
- - Select -

**Sort by:**
- Level
- Ascending
- 20

<table>
<thead>
<tr>
<th>Level</th>
<th>Can-do statement</th>
<th>Example</th>
<th>Details</th>
</tr>
</thead>
</table>
| A1    | **FORM:** COMBINING TWO ADJECTIVES WITH 'AND'
Can use 'and' to join a limited range of common adjectives. | Example | Details |
| A1    | **FORM:** WITH 'VERY'
Can use 'very' with a limited range of common gradable adjectives. | Example | Details |
| A1    | **FORM:** ATTRIBUTIVE (WITH NOUNS)
Can use a limited range of adjectives attributively, before a noun: noun phrases | Example | Details |
| A1    | **FORM:** PREDICATIVE, WITH 'BE'
Can use a limited range of adjectives predicatively, before a noun: noun phrases | Example | Details |

---

Paula Buttery
The English Grammar Profile
English Grammar Profile: What grammar?

- Select -
  adjectives
  adverb phrases - form
  adverbs
  adverbs and adverb phrases: types and meanings
  adverbs as modifiers
  alternatives
  articles
  can
  combining
  comparatives
  conditional
  coordinated
  coordinating
  could
  dare
  declarative
  demonstratives
  discourse markers in writing
  expressions with be
  focus
  future continuous
  future expressions with be
  future in the past
  future perfect continuous
  future perfect simple
  future simple (with will and shall)

Paula Buttery
The English Grammar Profile
<table>
<thead>
<tr>
<th>SuperCategory</th>
<th>SubCategory</th>
<th>Level</th>
<th>Can-do statement</th>
<th>Example</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODALITY</td>
<td>may</td>
<td>C2</td>
<td><strong>USE:</strong> CONCESSION Can use 'may' in a subordinate clause expressing concession with 'however', 'whatever', 'whoever' + 'it or this may be or seem'.</td>
<td>Example</td>
<td>Details</td>
</tr>
<tr>
<td>MODALITY</td>
<td>may</td>
<td>C2</td>
<td><strong>USE:</strong> HEDGING Can use 'it may be' + '-ed' that as a hedging device when drawing conclusions and summarising, particularly in academic contexts.</td>
<td>Example</td>
<td>Details</td>
</tr>
<tr>
<td>MODALITY</td>
<td>may</td>
<td>C2</td>
<td><strong>FORM/USE:</strong> 'MAY AS WELL' Can use 'may as well' to talk about what we think is the easiest or most logical course of action, often when we cannot see a better alternative or when it will not spoil the situation if you do that thing. ★ might</td>
<td>Example</td>
<td>Details</td>
</tr>
<tr>
<td>MODALITY</td>
<td>may</td>
<td>C1</td>
<td><strong>FORM/USE:</strong> 'MAY WELL' Can use 'may well' to give emphasis to something unexpected.</td>
<td>Example</td>
<td>Details</td>
</tr>
<tr>
<td>MODALITY</td>
<td>may</td>
<td>C1</td>
<td><strong>FORM/USE:</strong> PAST NEGATIVE Can use 'may not have' + '-ed' to talk about possibility in the past.</td>
<td>Example</td>
<td>Details</td>
</tr>
</tbody>
</table>
Have we learnt any lessons in construction and validation?

Questions we might ask:

- Can we build resources like the EGP automatically?
- How can we validate resources like the EGP?
Texts can be annotated using NLP techniques

"Project researchers use a statistical parsing tool"

CREATE PARSE TREES

EXTRACT GRAMMATICAL RELATIONS

NP

VP

VV0

AT1

AP

NN1

NN2

Subj

DetMod

Dobj

NounMod

NounMod

NounMod

NounMod

NounMod

NounMod

NounMod

NounMod

NounMod

NounMod

NounMod

NounMod

NounMod
Texts can be annotated using NLP techniques

(Project:1_NN1 researcher+s:2_NN2 use:3_VV0 a:4_AT1 statistical:5_JJ parsing:6_NN1 tool:7_NN1)

(ncsubj use:3_VV0 researcher+s:2_NN2)
(dobj use:3_VV0 tool:7_NN1)
(det tool:7_NN1 a:4_AT1)
(ncmod tool:7_NN1 statistical:5_JJ)
(ncmod tool:7_NN1 parsing:6_NN1)
(ncmod researcher+s:2_NN2 Project:1_NN1)
Structure can be extracted from the annotated text

Verbs can be automatically analysed according to the types of complement they take.

**Verb Constructions from Grammatical Relations**

he loves her

Will be represented with grammatical relations:

(\text{SUBJ} loves he)

(\text{OBJ} loves her)

TRANSITIVE or NP
Structure can be extracted from the annotated text

Verb Constructions from Grammatical Relations

\[(\text{SUBJ} \ loves:VVZ \ he:PPHS1)\]
\[(\text{OBJ} \ loves:VVZ \ her:PPH01)\]

TRANSITIVE or NP

Rule based classification

- **if** output GR patterns match
  \[(\text{SUBJ} \times y)\]
  \[(\text{OBJ} \times z)\]
  where \(x, y\) and \(z\) are variables
- **and** the tag attached to \(x\) starts with "V"
- **then** classify verb at \(x\) to be NP
Structure can be extracted from the annotated text

We can now search for constructions in the annotated texts...

- 168 distinct verb patterns (a superset on the COMLEX (Grishman et. al, 1994) and ANLT (Boguraev et al., 1987) syntax dictionaries)
- Adjectival patterns
- Compound noun patterns
- Relative clause patterns
- ... and any other patterns you can specify

We can search for any constructions we like ...

... as long as we know what to look for!
EVP and EGP demonstrate criterial features

The goal of EGP is to capture the grammar that learners are using appropriately and correctly at each CEFR level.

[Criterial Features] are properties that are distinctive and characteristic of each of the levels and hence important for both practitioners and theoreticians.

The first set of practitioners who can benefit from the identification and explanation of criterial features are the learners of English themselves and those who teach them.

Hawkins and Filipovic, 2012
The grammar of English is complex but finite.

Researchers drew up a long list of grammar from:
- Reference books e.g. Cambridge Grammar of English
- Text books e.g. the Cambridge course books
- Constructions from the Cambridge English Corpus

Example constructions associated with this grammar list can be found in the Learner Corpus by exploiting the NLP annotations as described.

How do we know we found all the relevant grammar?
Can we automate the process of drawing up this list?
Machine Learning techniques provide holistic classification

Brief intro to Machine Learning:

A machine learning classifier can automatically assigned CEFR levels to previously unseen texts based on the holistic properties (features) of the text.

Successful machine learning requires:

- Methods for extracting features
- A very large amount of representative training data
We can represent learner language as predictive features.
Then some though occurred to me

1. Word sequences
   - (unigram) Then, some, though, occurred, ...
   - (bigram) Then some, some though, though occurred, ...
   - (trigram) Then some though, some though occurred, though occurred to, ...

   Then_RR some_DD though_RR occurred_VVN to_Il me_PPI01

2. Linguistic category sequences
   - RR DD (e.g. Then some)
   - RR VVN (e.g. occurred to)
Grammatical constructions using Robust Accurate Statistical Parser

- NP/a1-cat_np-r (e.g. *Then some though occurred*)
- PP/p1 (e.g. *to me*)
Then some thought occurred to me

Grammatical relation constructions:
(subj occur_V N) and (iobj occur_V to) and (dobj to PP)

The features have been listed here in order of decreasing frequency in the text... frequencies are different by orders of magnitude
## Discriminative features for predicting a CEFR level

<table>
<thead>
<tr>
<th>Feature</th>
<th>Weight</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP/a1</td>
<td>5.99959</td>
<td>very cleverly</td>
</tr>
<tr>
<td>Apostrophes</td>
<td>3.60846</td>
<td>yesterday’s weather</td>
</tr>
<tr>
<td>PP/p1</td>
<td>3.60492</td>
<td>in the mornings</td>
</tr>
<tr>
<td>NP/det_n1</td>
<td>2.26438</td>
<td>the film</td>
</tr>
<tr>
<td>would</td>
<td>1.65683</td>
<td>would rather go</td>
</tr>
<tr>
<td>T/frag</td>
<td>-1.94559</td>
<td>but know she knew</td>
</tr>
<tr>
<td>cas</td>
<td>-1.21774</td>
<td>occasion</td>
</tr>
<tr>
<td>VBDZ_RR</td>
<td>1.19119</td>
<td>was truly demanding</td>
</tr>
<tr>
<td>VB0_VVN</td>
<td>1.14131</td>
<td>be involved in</td>
</tr>
<tr>
<td>-ily</td>
<td>-1.13626</td>
<td>easily</td>
</tr>
<tr>
<td>PPIS2_VM_VV0</td>
<td>-0.939866</td>
<td>as we can see</td>
</tr>
</tbody>
</table>
Discriminative features are not criteria features

- Hundreds of thousands of low-level discriminative features
- Discriminative features are proxies to aspects of the criterial features and need interpretation

![Graph showing the contribution of feature types]

**Figure 3: Contribution of Feature Types**

<table>
<thead>
<tr>
<th>Feature Types</th>
<th>CLC Rater 1</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Auto-mark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terms + bigrams + corpus err. rate + parse rules + all other features</td>
<td>71.9%</td>
<td>73.4%</td>
<td>73.8%</td>
<td>76.7%</td>
<td>76.7%</td>
<td>76.7%</td>
</tr>
<tr>
<td>Terms + bigrams + corpus err. rate</td>
<td>69.0%</td>
<td>70.5%</td>
<td>71.9%</td>
<td>73.4%</td>
<td>73.8%</td>
<td>76.7%</td>
</tr>
<tr>
<td>Terms + bigrams + corpus err. rate</td>
<td>66.4%</td>
<td>68.9%</td>
<td>71.9%</td>
<td>73.4%</td>
<td>73.8%</td>
<td>76.7%</td>
</tr>
</tbody>
</table>

**Table 2: Correlation (Pearson’s CC)**

We also compared the performance of the preference ranking TAP model to a binary TAP classifier trained using the same feature types on the same data divided into pass/fail scripts. The correlation with the CLC scores on this test data was worse by 0.05 (Pearson) and 0.07 (Spearman) using classification as compared to preference ranking with the same underlying TAP model.
Goals of the EPP
Constructing the EGP
Validation of EGP
Annotating text using NLP
Criterial features
Discriminative features

The Discriminative Feature Visualiser

Graph properties:
- Discriminative power
- Co-occurrence threshold: 0.5 - 1
- Total matching nodes: 6

Feature-Error relations:
- Features (pos:2325, neg:10)
  1. VM_RR (+)
  2. VM_RR_VV0 (+)
  3. _because (-)
  4. NN1_VV0 (-)
  5. how_to (-)
  6. the_people (-)
    weight: 2.62449

Errors
- 27.9: UD (unnecessary)
- 20.95: S (spelling error)
- 18.35: RT (replace present)
- 17.79: MP (missing parameters)
- 14.65: RV (replace verb)
- 14.27: RP (replace present)
- 13.72: TV (incorrect verb)
- 11.03: AGV (verb agreement error)
- 10.1: R (replace error)
- 10.1: RN (replace noun)
- 9.18: MD (missing data)

Search:
- Search corrected text
- Sentence-level search
- Tokens inside errors
- Search corrected tokens
- Tokens near errors
- Text contains errors
- N-grams preceded/followed by errors

Output:
- orig_error-coded text
- GRS (grammatical relations)
- Include meta-data
- Separate by grade

Number of hits: 500

Paula Buttery
The English Grammar Profile
### The Discriminative Feature Visualiser

**FCE scripts**

<table>
<thead>
<tr>
<th>grade: A</th>
<th>grade: B</th>
<th>grade: C</th>
<th>grade: D</th>
<th>grade: E</th>
</tr>
</thead>
</table>

**Errors by decreasing frequency**

Display errors:

- **Display options**: Save, search results: how to, Go, <, >, 19 matches found.

**Searching for**: +TX_inc_text: "how to" +TX_inc_pots-seq: "RGQ TO VVO"

Searching in index: FCE_v4_grades/FCE_v4_gradeC containing 2004 learners.

191 total matching sentences

**Example Sentences**:

1. **22820.0**: As for the camera, I would like to have one but I don't know **how to use it**.

   <error-coded-text>
   22820.0 As for the camera, I would like to have one but I don't know **how to use it**.
   </error-coded-text>

2. **22820.1**: As for the camera I would like to have one but, I don't know **how to use it**.

   <error-coded-text>
   22820.1 As for the camera I would like to have one but, I don't know **how to use it**.
   </error-coded-text>

3. **44248.0**: However I do have some suggestions **how to make it even better**.

   <error-coded-text>
   44248.0 However I do have some suggestions **how to make it even better**.
   </error-coded-text>

4. **44248.1**: However I do have some suggestions **how to make it even better**

   <error-coded-text>
   44248.1 However I do have some suggestions **how to make it even better**.
   </error-coded-text>

5. **98608.0**: This is a **really, really** good suggestion, but before being considering, **how to realise that**

   <error-coded-text>
   98608.0 This is a **really, really** good suggestion, but before being considering **how to realise that**
   </error-coded-text>

**Total word count**: 66411
Hypothesis generation from discriminative features

18th most discriminative (negative) feature (B1/B2 level)

- **RG_JJ_NN1**: Degree adverb followed by an adjective and a singular noun (e.g., *very good boy*)
- Why negative?
- Co-occurs with:
  - `very good` (-)
  - `JJ_NN1_Il` (-) (e.g., *difficult sport at*)
  - `VBZ_RG` (-) (e.g., *is very*)

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

1b) *I think it’s very good idea to spending vacation together.*
Hypothesis generation from discriminative features

With complex adjectival phrases learners may omit the article:

**Hypothesis for B2 criterial feature**

By B2 learners can appropriately and correctly *use degree modifiers followed by an adjective describing a single noun* (e.g. a very good boy).
Hypothesis generation from discriminative features

Analysis shows this primarily the case for learners from L1s lacking articles:

<table>
<thead>
<tr>
<th>Language</th>
<th>RG_JJ_NN1</th>
<th>VBZ_RG_JJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>2.75</td>
<td>2.73</td>
</tr>
<tr>
<td>Turkish</td>
<td>5.81</td>
<td>5.82</td>
</tr>
<tr>
<td>Japanese</td>
<td>4.48</td>
<td>3.98</td>
</tr>
<tr>
<td>Korean</td>
<td>5.48</td>
<td>5.31</td>
</tr>
<tr>
<td>Russian</td>
<td>5.42</td>
<td>4.59</td>
</tr>
<tr>
<td>Chinese</td>
<td>3.58</td>
<td>3.25</td>
</tr>
<tr>
<td>French</td>
<td>1.32</td>
<td>1.49</td>
</tr>
<tr>
<td>German</td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td>Spanish</td>
<td>1.18</td>
<td>1.35</td>
</tr>
<tr>
<td>Greek</td>
<td>1.60</td>
<td>1.70</td>
</tr>
</tbody>
</table>
Goals of the EPP
Constructing the EGP
Validation of EGP

Validation of the EGP using discriminative features

SuperCategory: Modality; SubCategory: must

B2: past affirmative form:
  - Can-do: Can use must have + -ed.
  - Example: When I realized that he must have forgotten, I called him up.

B2: use to express deductions and conclusions:
  - Can-do: Can use the perfect form of must to make deductions about the past.
  - Example: I must have looked awful because she asked me whether I was all right.
**Validation of the EGP using discriminative features**

*Must* is a highly-ranked negative* discriminative feature related to:

- **VM_VVD** (negative): modal auxiliary_paste tense
- **VM_TO** (negative): modal auxiliary_{to}
- **must_be** (negative)

*Negative ⇒ autograder evidence that the learner is not at B2 level yet, i.e., has not mastered the rule related to the feature
Conclusions

- Discriminative features are highly useful for holistic classification tasks (such as auto-marking) but they are not interpretable (criterial features)
- It is not straightforward to generate criterial features from discriminative features
- The top-down and bottom-up approaches of criterial and discriminative features may corroborate each other through specific case studies
- Discriminative features may provide hypotheses for future criterial features
- The more mappings we have between discriminative and criterial features the better we can predict criterial features in the future