Automatic EFL Proficiency Assessment via detailed and deep feature extraction

Mick O'Donnell
Universidad Autónoma de Madrid

Aim

- An experiment to see to what extent automatically annotated learner texts can be used to **predict the learner's grammatical proficiency level**.

  (Inspired by the talk in last year's CILC by María Ángeles Zarco-Tejada)

Proficiency

- There are various types of learner proficiency (oral, written, listening, and in writing, vocabulary, grammatical, discoursal, organisational, etc.)
- We are focusing here on grammatical, and thus ‘use of english’ proficiency.
- Each learner in our study was graded for proficiency using the Oxford Quick Placement Test (60 questions, use of English)

Prior Work

- Massive amount of work in this area, much on automatic oral assessment, not relevant here.
- Work on written assessment often uses **lexical clues (**word frequency, sentence length, lexical diversity, word repetition, text length, …) e.g, Reid, 1986; Connor, 1990; Reppen, 1994; Ferris, 1994; Jarvis 2002 etc.
- More recent work using automatically derived **syntactic features** (e.g., Scott et al, 2014)
- Others use some **discourse patterns** (e.g., cohesion) or **rhetorical features** (argumentation; Attali, 2007)

Methodology

1. Automatically annotate a large number of learner texts for **lexical, syntactic and discourse-semantic features**.

2. Identify **level of use** of each feature in each text.

3. Associate **proficiency level** (0-60) with each essay from placement test. (Oxford Quick PT)

4. **Build statistical model** to predict proficiency given levels of linguistic patterns.

Methodology

- NOTE: most other work uses human assessment of quality of essay as input, and then looks for factors in the text which correlate with high/low scores.
- Here, the measure of proficiency is **external** to the text
- But we assume ability in a placement test should correlate with patterns in their linguistic production.
- We are trying to locate those aspects of learner writing that most reflect grammatical proficiency.

Corpus

- **WriCLE Corpus** (Rollinson & Mendikoetxea, 2010)
  - 556 essays by Spanish University learners of English (approx. 1725 words each) each with associated proficiency score.
- 74 **BAWE** Sociology Essays (similar questions by English natives)

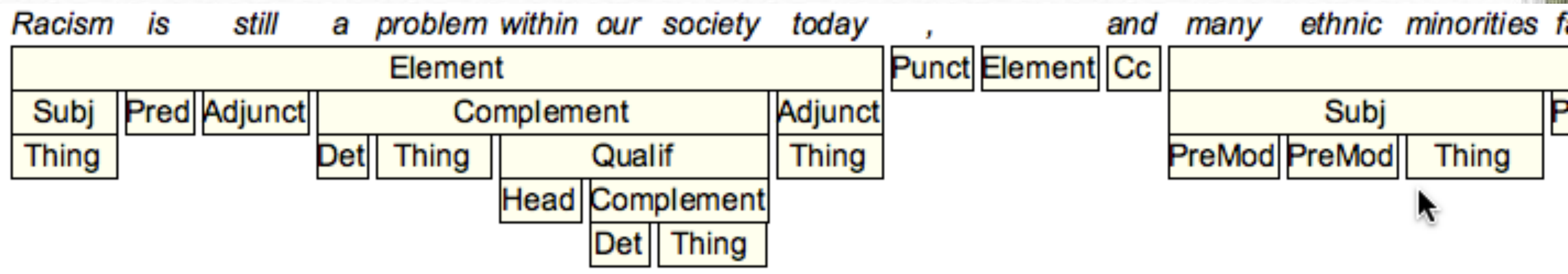
Linguistic Annotation (i)

- **General lexical statistics:**
  - Average word length
  - Average sentence length
  - Pronoun use (1stPersSing, 1stPersPlur, 2ndPers, 3Pers)
  - Lexical density (lexical words % of all words)
  - Subjective positivity (ratio of +ve to -ve words)

Grammatical Annotation

- Automatic Syntactic Annotation by Stanford parser within UAM Corpustool
- Transformed into more semantic form by UAMCT (transitivity, theme-rheme, modality, etc.)

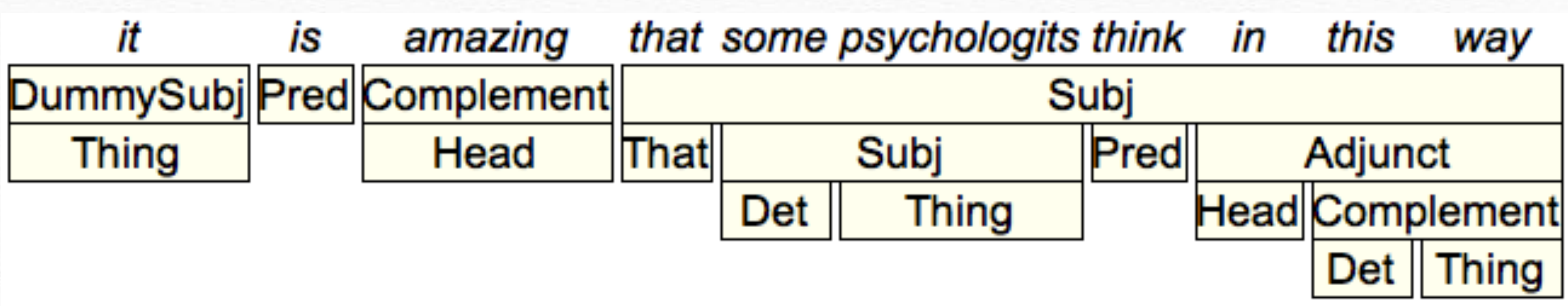
Basic Grammar



- Clause Features:
  - **Voice** (active vs passive)
  - **Tense-Aspect** (simple-present, past-perfect, etc.)
  - **Mood** (declarative, interrogative, imperative)
  - **Finiteness** (finite, infinitive-clause, past-participle-clause, present-participle-clause, relative-clause, that-clause, etc.)
  - **Marked Sentence Structure**: it-cleft, extraposition, there-existential, etc.

Featurisation

- The parser produces a functional role (e.g., Subj) and one class feature for each constituent.
- To be useful for this kind of study, we need to **featurise** the data:
  - recognition of structural patterns and adding a tag for this.



‘it’ + [be] +comment-adj +that-clause -> extraposition

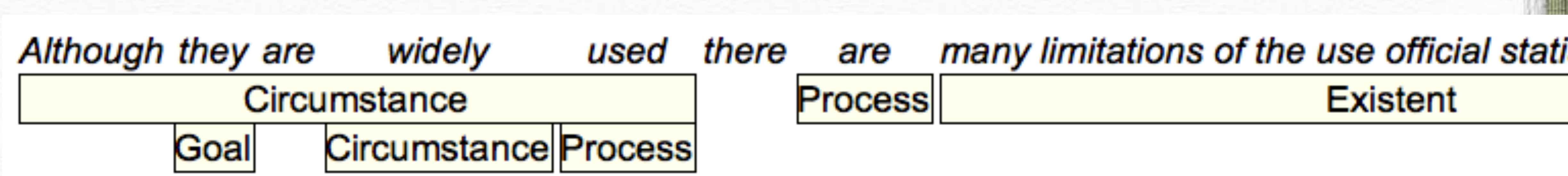
Modality

- **Syntactic types**
  - modal auxiliary, (*should*)
  - semi-lexical (*have to, ought to*),
  - verb (*require*),
  - adverb (*possibly*)
  - adjective (*it is possible*)
- **Semantic types (of lexical modals)**
  - possibility, necessity, obligation, etc.

(based on work with Rebeca Garcia)

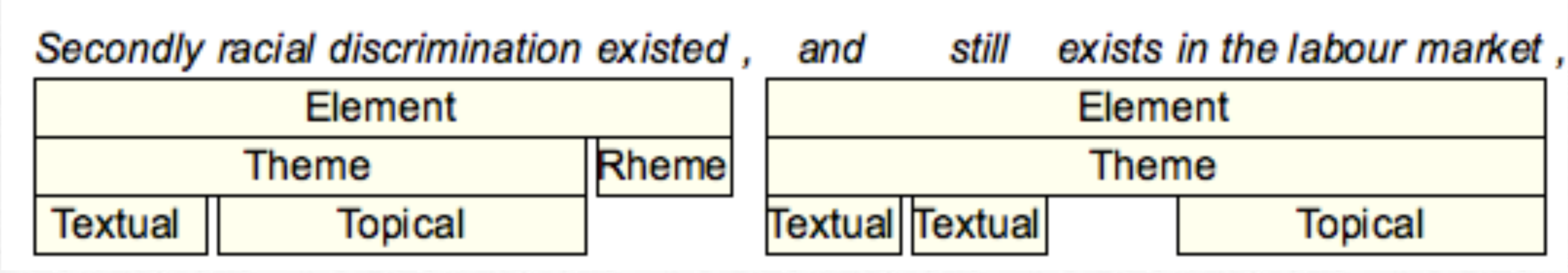
Transitivity

- **Recognition of semantic roles**
  - Actor, Process, Goal, Sensor, Phenomenon, etc.
- **Each clause assigned a process type**
  - material, mental, verbal, relational, existential
- **Key patterns recognised:**
- **verbal-passive** (*it has been said that…*)
- **mental-passive** (*it is believed that…*)
- **Say-type vs. tell-type,**
- **please-type vs. like-type**



Theme-Rheme

- Recognition of Topical, Interpersonal and Textual Themes (Halliday)
  - **Textual**: conjoin clause to previous clauses.
  - **Interpersonal**: Speaker comment or provision of probability etc. (*Luckily, apparently*, etc.)
  - **Topical**: The first ideational item in the clause

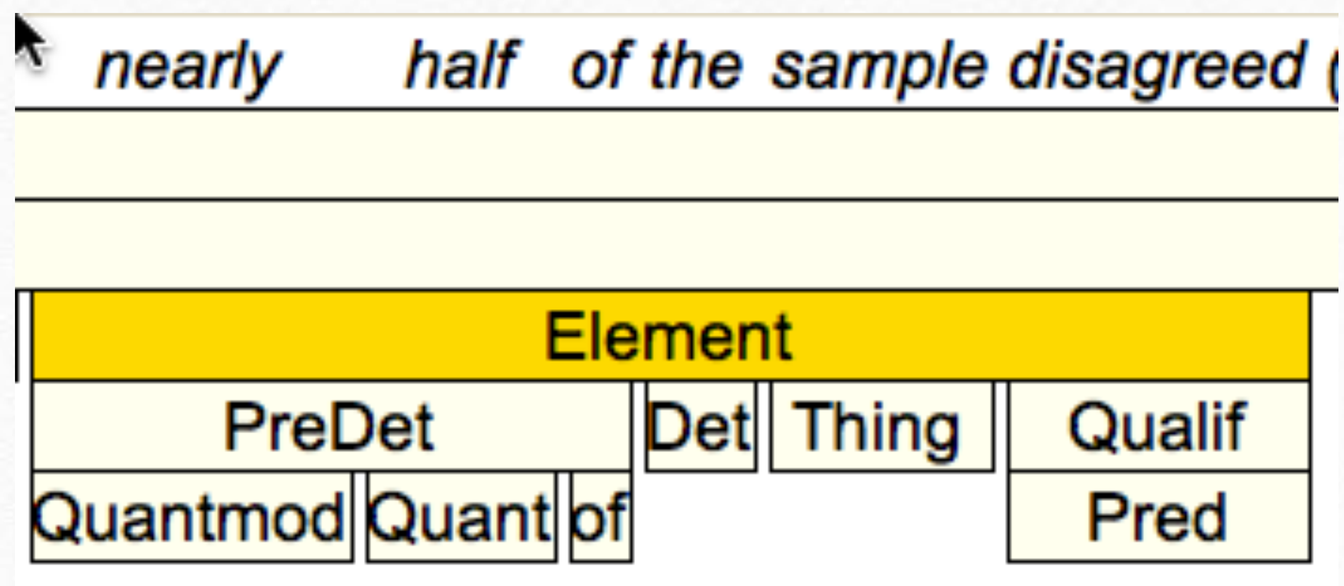


Theme-Rheme

- Featurised in terms of:
  - **degree** of use of textual, interpersonal themes
  - **marked** topical themes: *fronted-adjunct*, *elided-theme*, *dummy-theme*, etc.
  - **textual** semantic types: *structuring* (*firstly*), *arguing* (thus), *extending* (and)
  - **interpersonal** semantic types: *evidence (probably)*, *evaluation (happily)*, *admission (honestly)*, etc.

Noun Phrase

- **Noun Phrase Structure**:
  - **predetermined** (*all the children, all of the children*)
  - **determiner type** (none, the, many, another, etc.)
  - **premodification** / **postmodification**
  - **Kind**: proper, common, pronominal
  - **Extensive quantification features**
  - **count vs. mass nouns**
  - **abstract vs. concrete nouns**
  - **nominalised heads** (*the run, the dismissal*, etc.)



Data Summary

- 250+ linguistic features pruned back to the 170 most likely to reflect proficiency.
- 630 essays fully annotated
- Levels of use of each feature extracted to a spreadsheet.
- 50 **testing files** split off into a reserve.
- 580 files in the **training set**.

Linguistic Modeling

- First experiment: **multiple regression**
  - Profic.= $a.F_1 + b.F_2 + c.F_3 + \ldots..$
  - Used a hillclimbing method to find best values of a, b, c, etc. to maximise accuracy of predicting proficiency of the training set.
  - Then applied this model to the test set…

Hill Climbing Multitple Regression

- All parameters initially set to 0
- On each iteration, test changes (+/- 0.01) to each parameter to produce the formula
- For each change, measure differences between predicted proficiency and test score.
- Keep change with smallest sum of square difference.

Iterative solutions

- P = -0.5*modal-auxilliary +52.39
- P = -0.5*modal-auxilliary **-0.5*3pRef** +**54.16**
- P = -0.5*modal-auxilliary -**1***3pRef + **55.92**
- P = -0.5*modal-auxilliary -1*3pRef + **0.5*AvWdLen** + 53.55
- P = -0.5*modal-auxilliary -1*3pRef + **1.0***AvWdLen + 51.18
- P = -0.5*modal-auxilliary -1*3pRef + **1.5***AvWdLen + 48.81
- etc.

Final solutions: Positive factors

- Supporting high proficiency: (bigger numbers mean bigger impact)
  - qualified-group 29.0 (postmodif. in noun phrase)
  - passive-clause 17.5
  - nonfinite-clause 13.0
  - abstract-noun12.0
  - interrogative-clause 10.0 (rhetorical questions)
  - arguing 9.0 (thus, in consequence, etc.)
  - no-quantifier-agreement-error 8.5
  - improbability 8.0 5.5 (*it is unlikely*…)
  - most-determined 7.5 “most people”
  - not-determined-group 7.0 (*people*)
  - elided-ideat-theme 7.0 “*and believed that*”
  - exclamative-predetermined 7.0 (*such a situation*)
  - Fronted-adjunct 6.5 “In 1865, …

Final solutions: Negative factors

- Supporting low proficiency: (bigger numbers mean bigger impact)
  - summative -5.5 "in summary"
  - each-determined -7.5 "each person"
  - enough-determined -10.5 "enough problems"
  - simple-present -11.0
  - present-progressive -16.5
  - 1p-plur-19.0 "I believe"
  - plural-noun -11.5

Overall Results

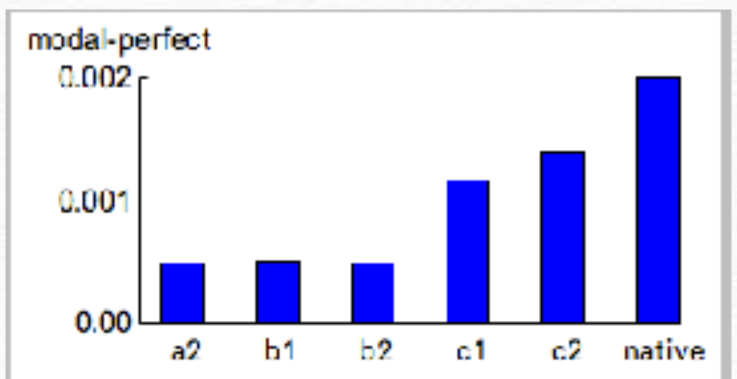
- Pearson **correlation** coefficient of **0.68** (correlating predicted proficiency with actual proficiency over 50 text test set)
- **Average error** in prediction **6.2** (out of 60)
- Lower than many systems which assign grades to essays
- But we are not grading the essay but the use of english proficiency
- Many are commercial systems with lots of fine tuning
- I have not built in many of the lexical factors which correlate most highly with proficiency (academic word level, type-token ratios, etc.)
- Parsing of learner texts less reliable than native texts, thus higher error rate in some usage levels.
- Scope for improvement:Some Syntactic analyses < 90% accurate (it-cleft, ditransitive-verb, imperative, etc.)

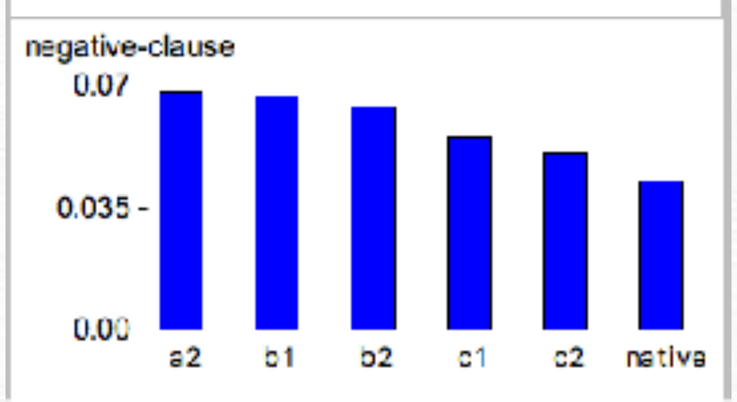
Scope for improvement

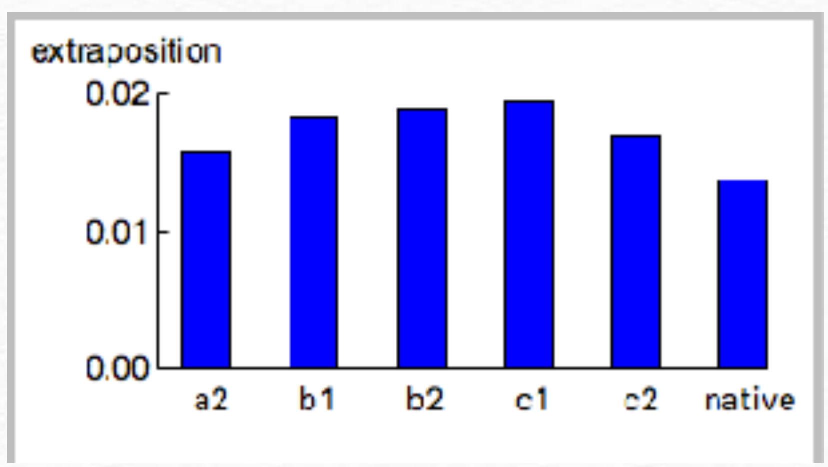
- Some Syntactic analyses < 90% accurate (it-cleft, ditransitive-verb, imperative, etc.) and I can improve this.
- More data in will give better results.
- I have not normalised the usage levels, which may improve the results.

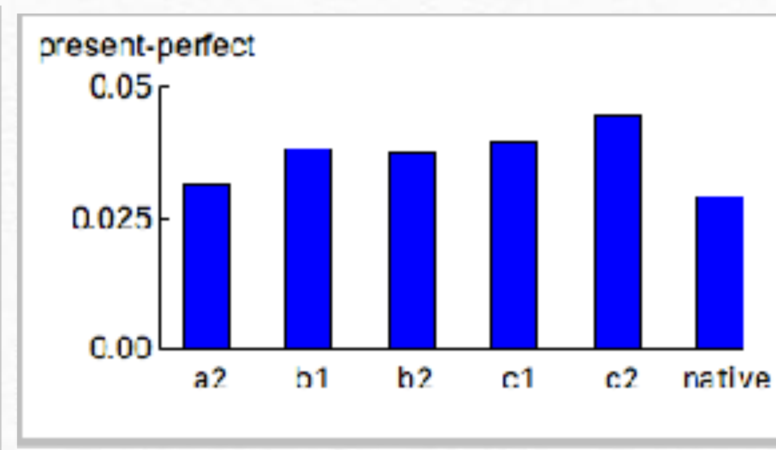
Problem

- Some variables are not clearly correlated with proficiency, possibly because of rising/falling acquisition patterns
- It may be the case that some factors are more important indicators at different proficiency levels, or indicative levels differ for lower, intermediate and advanced learners.
- In some cases, clear patterns in the learner levels contradicted by native data.









Solution: Prototype clustering

- The program searches for prototype learner profiles which best explain the patterns in the data.
- We initially set the number of prototype profiles to use (e.g, 6)
- Each document associated to the prototype it is most similar to (a cluster)
- System tests each possible mutation of each prototype (increase factor, decrease factor),
- Documents are then reassigned.
- The mutation that gives the best clusterings of documents in terms of similarity of proficiency is kept.

Solution: prototype clustering

- Process produces 6 prototype profiles, which cover the space from beginner to advanced to native.
- Prototypes only allowed to include 12 factors at most.
- Overall predictivity not so good
  - Correlation with actual proficiency in test set: 0.57
  - Average error: 6.47
  - BUT interesting groupings

Solution: prototype clustering

- Highest model: Average proficiency **60.57** (the native texts were assigned a proficiency score of 62 by default) - so, nearly all native and some high learner texts.
- Factors:
  - Av. sentence Length: 25.93 words
  - Av. Word Length:: 4.99 characters
  - 3p pronouns: 26.2 tokens per 1000 words.
  - extraposition: 1.24% of clauses
  - verbal-process: 5.2% of clauses
  - past-tense: 32.4% of finite clauses
  - post modified NP: 33.2% of noun-phrases
  - elided-ideat-theme: 3.4% of clauses
  - demonstrative-determined: 7.6% of noun phrases
  - extending connectors (and, etc.) : 9.2% of connectors

Discussion

- This prototype-based clustering technique is interesting because it allows for distinct types of learners to be identified and separated.
- Learners with similar test scores may reflect different language backgrounds
    - E.g., natives vs high level Spanish learners
    - E.g., quick learner with no experience vs. long term learner who is bad at language.
- However, at present I haven't found the right way to configure the models to make the hillclimbing search work optimally.
- Tends to produce several groups in the centre, rather than spread out over the levels.
- Lots of variables to handle.

Conclusions

- This paper has discussed two experiments in the use of a large linguistically annotated corpus to build models which can be used to predict use of grammatical proficiency.
- A corpus of 580 training essays,
- Over 170 distinct linguistic features automatically tagged.
- Multiple Regression model produced ok results (0.68) but not up to commercial levels.
- But more work on refining linguistic accuracy and introducing more relevant factors may help this.

V

- The prototype version produces interesting results, but even less accurate.
- But good indicator of which features are important at different levels.
- I will continue to refine the search mechanism to produce better clustering of documents matching proficiency types.