Colored Organized Lexical Recognition Software

Cory Morewitz, Sean DeMild

Justin Bennet, Erich O'Saben

James Ward, Charles Stafford

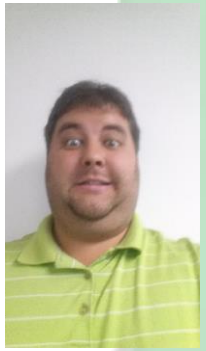
Christopher Hultin

Mentor:
Professor Greg Raver-Lampman

Outline

Meet the Team

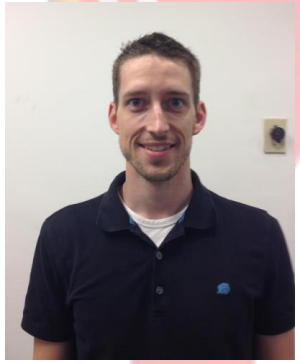
Professor Greg Raver-Lampman – Mentor



Cory
Project
Manager

Sean
Lead
Designer

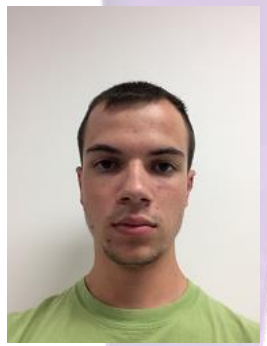
Justin
System
Engineer



Erich
Lead
Python
Developer

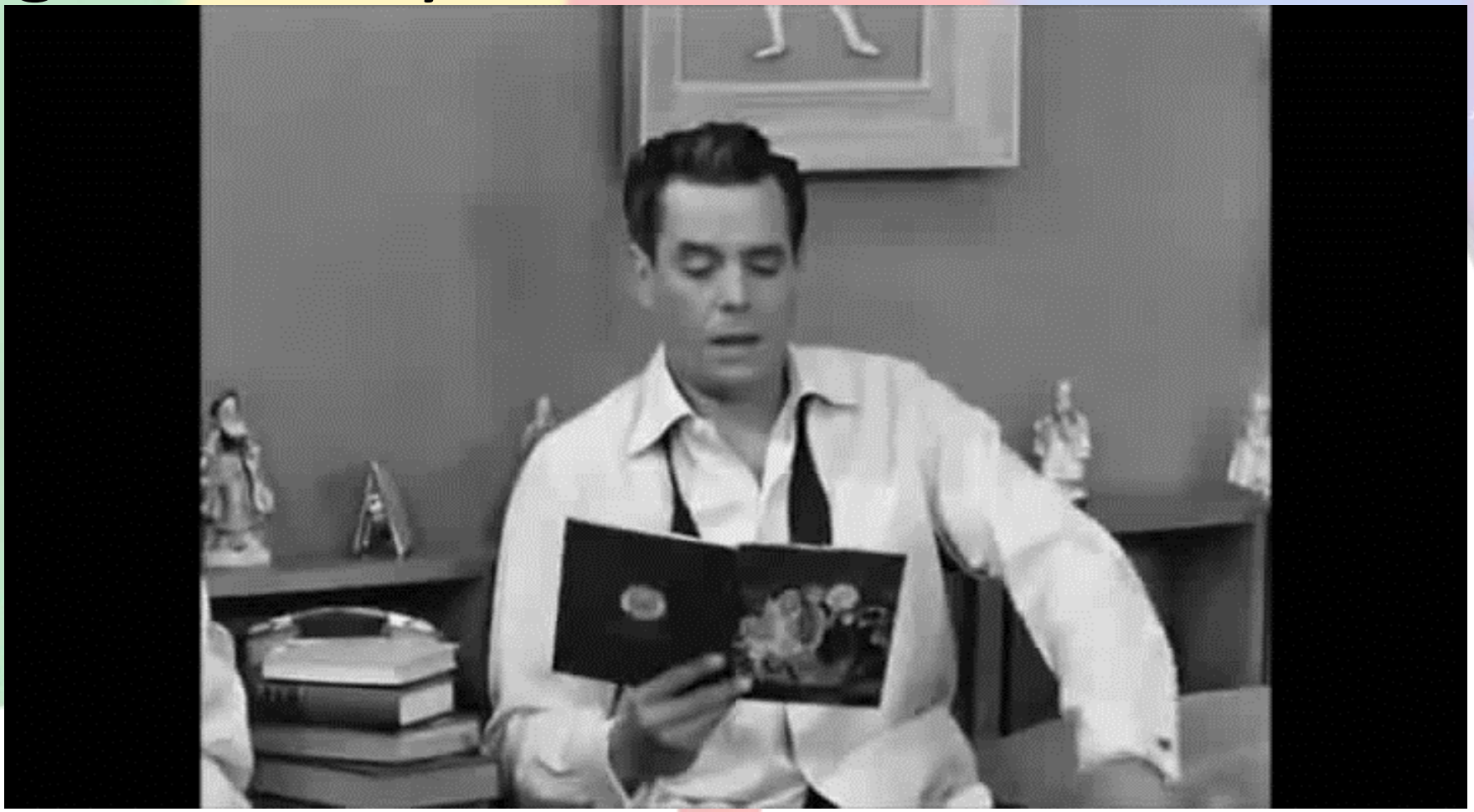
Charles
Python
and
Risk
Management

James
System
Architect



Christopher
Website
and
System
Admin

Is English Really That Difficult?

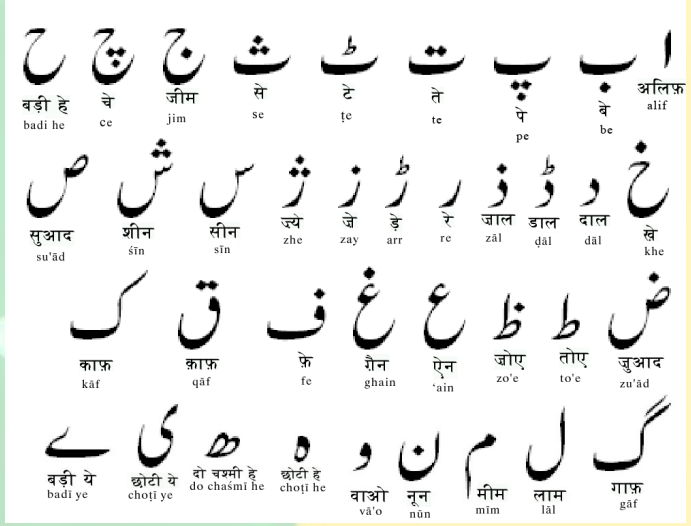


English Is Hard

- "Ricky" can speak and read English, but has difficulty pronouncing similarly spelled words
- In languages such as Spanish, distinct letter combinations are pronounced the same way every time
- This can make memorizing the different quirks of English far more troublesome for ESL students whose first languages have set rules
- Unfortunately varying pronunciations are not the only problem ESL students face

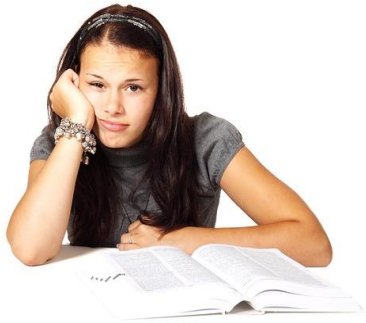
Problem Statement

ESL Instructors have observed a shared difficulty among students in identifying parts of speech in the English language due to the highly variable sentence structures in the language.



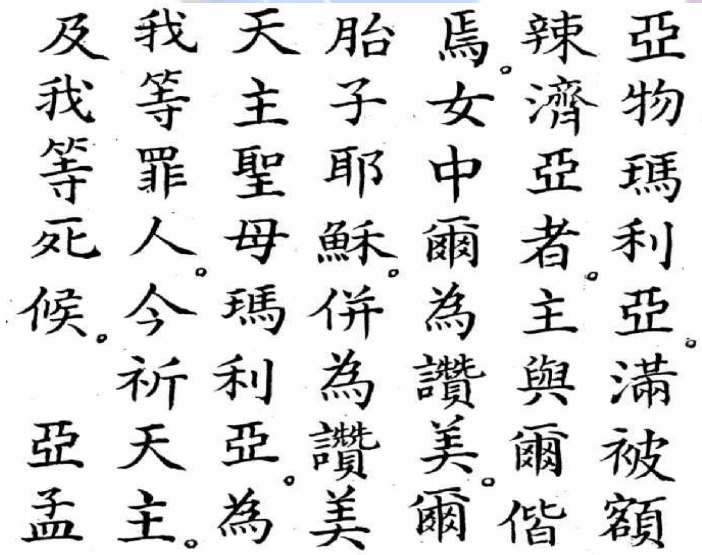
Case Study

●Foreign students in ESL

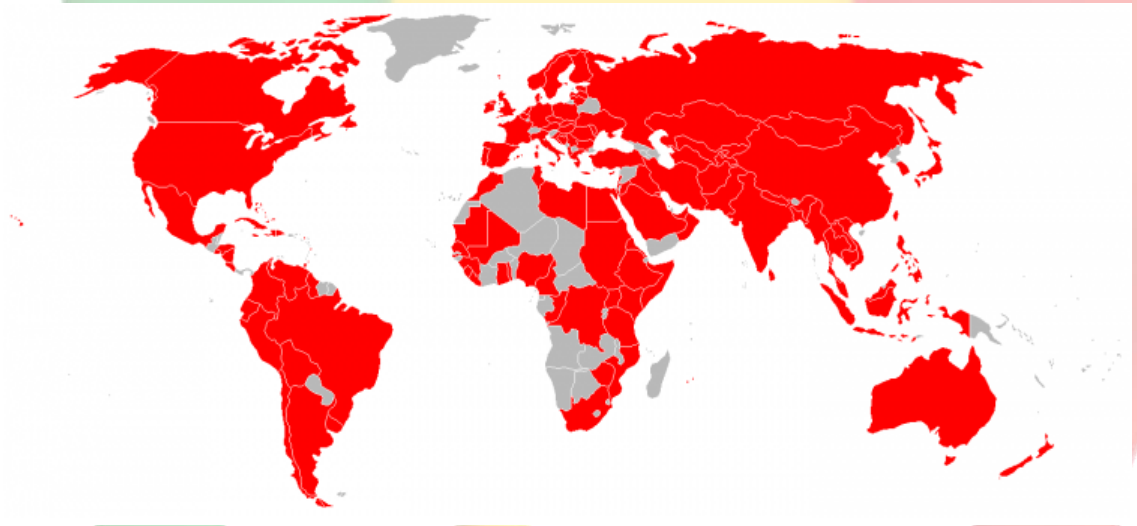


● Reading most difficult part of learning English

● How can P.O.S. be identified to facilitate learning



Countries Represented by ODU Students



- Over 200 international students per year
- 90% success rate
- 1980 286 thousand students in ESL programs
- 2010 Greater than 690 thousand students

Potential Market

Table 282. Foreign (Nonimmigrant) Student Enrollment in College: 1980 to 2010

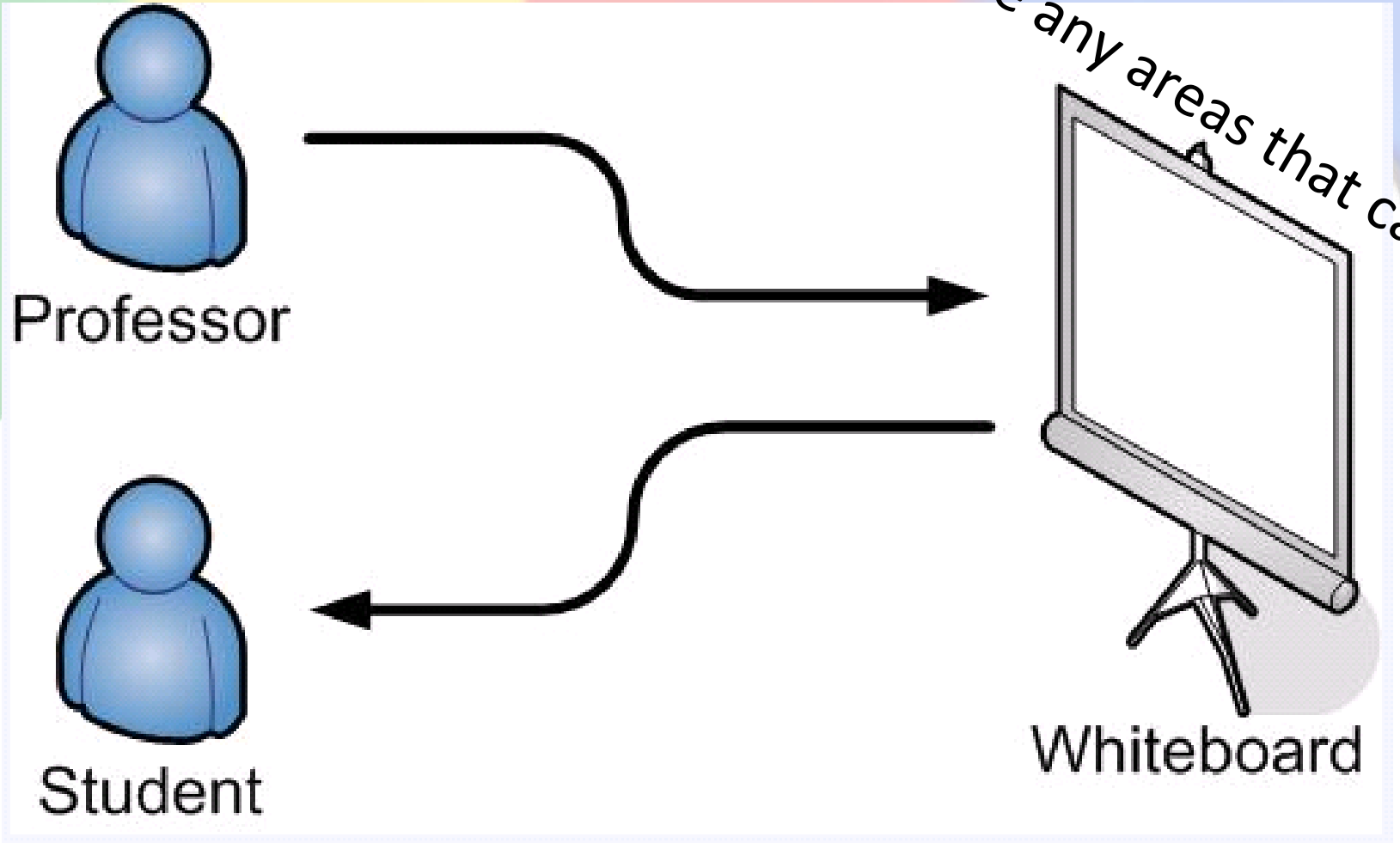
[In thousands (286 represents 286,000). For fall of the previous year]

Region of origin	1980	1990	1995	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
All regions	**286**	**387**	**453**	**491**	**515**	**548**	**583**	**586**	**573**	**565**	**565**	**583**	**624**	**672**	**691**
Africa	30	25	21	20	30	34	38	40	38	30	36	30	30	37	37
Nigeria	10	4	2	3	4	4	4	6	6	6	6	6	6	6	7
Asia [1,2]	165	245	292	303	315	339	353	367	356	356	346	367	405	444	469
China [3]	1	33	39	51	54	60	63	65	62	63	63	68	81	98	128
Taiwan [3]	18	31	36	31	29	29	29	28	26	26	28	29	29	28	27
Hong Kong	10	11	13	9	8	8	8	8	7	7	8	8	8	8	8
India	9	26	34	37	42	55	67	75	80	80	77	84	95	103	105
Indonesia	2	9	12	12	11	12	12	10	9	8	8	7	8	8	7
Iran	51	7	3	2	2	2	2	2	2	2	2	3	3	4	5
Japan	12	30	45	46	47	46	47	46	41	42	39	35	34	29	25
Malaysia	4	14	14	12	9	8	7	7	6	6	6	5	5	6	6
Saudi Arabia	10	4	4	5	5	5	6	4	4	3	3	8	10	13	16
South Korea	5	22	34	39	41	46	49	52	52	53	59	62	69	75	72
Thailand	7	7	11	12	11	11	12	10	9	9	9	9	9	9	9
Europe [4]	23	40	65	74	78	81	82	78	74	72	85	83	84	88	85
Latin America [4,5]	42	48	47	55	62	64	68	69	66	63	65	65	64	66	66
Mexico	6	7	9	10	11	11	13	13	13	13	14	14	15	15	13
Venezuela	10	3	4	5	5	5	6	5	6	5	5	5	4	5	5
North America	16	19	23	23	24	26	27	27	28	29	29	29	29	30	29
Canada	15	18	23	23	24	25	27	27	27	28	28	28	29	30	28
Oceania	4	4	4	4	5	5	5	5	5	4	5	4	5	5	5

Parts of Speech

POS	Function	Example
noun	Names a person,place of thing	Student,school,book
pronoun	Takes the place of a noun	I, you, she, he, it, them
verb	Identifies action or state of being	Run, jump, believe, be
adjective	Modifies a noun	Good, bad, ugly
adverb	Modifies a verb, adjective or other adverb	Quickly, easily, Always
preposition	Shows a relationship between a noun (or pronoun) and other words in a sentence	up, over, against, by, for
conjunction	Joins words, phrases, and clauses	and, but, or, yet
interjection	Expresses emotion	ah, whoops, ouch

Current Process

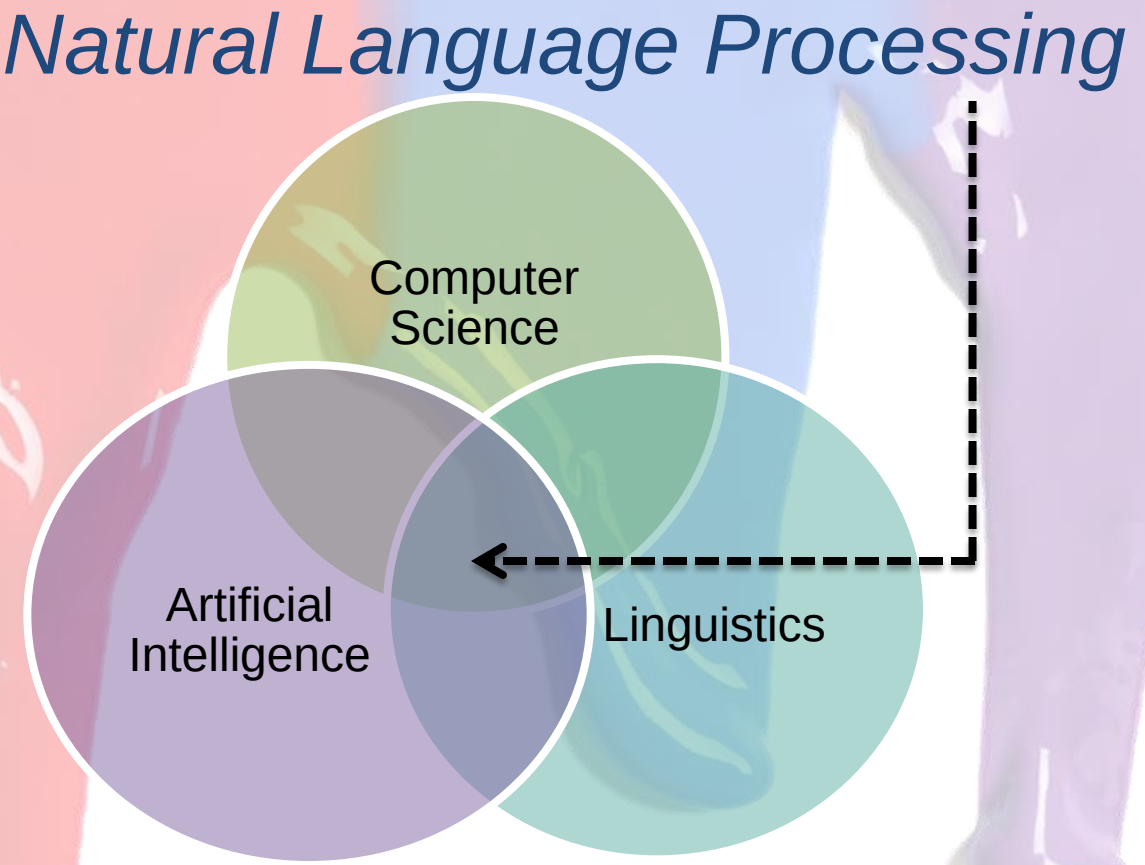


Can you see any areas that can be improved?

Natural Language Processing (NLP) And Parts of Speech Tagging

- NLP - Is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (**natural**) languages.

- Part of Speech Tagging - the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context



Examples of Current Solutions

- Part of Speech Taggers:

University of Illinois at Urbana

NNPS/ Helicopters MD/ will NN/ patrol DT/ the JJ/ temporary JJ/ no-fly NN/ zone IN/ around NNP/ New NNP/ Jersey POS/ 's NNP/ MetLife NNP/ Stadium NNP/ Sunday ,/ , IN/ with NNP/ F-16s VBN/ based IN/ in NNP/ Atlantic NNP/ City JJ/ ready TO/ to VB/ be VBN/ scrambled IN/ if DT/ an JJ/ unauthorized NN/ aircraft VBZ/ does VB/ enter DT/ the VBN/ restricted NN/ airspace ./ .

IN/ Down IN/ below ,/ , JJ/ bomb-sniffing NNS/ dogs MD/ will NN/ patrol DT/ the NNS/ trains CC/ and NNS/ buses WDT/ that VBP/ are VBN/ expected TO/ to VB/ take RB/ approximately CD/ 30,000 IN/ of DT/ the JJ/ 80,000-plus NNS/ spectators TO/ to NNP/ Sunday POS/ 's NNP/ Super NNP/ Bowl IN/ between DT/ the NNP/ Denver NNS/ Broncos CC/ and NNP/ Seattle NNP/ Seahawks ./ .

Xerox

- Helicopters
  - helicopter ***+NOUN***
- will
  - will ***+VAUX***
- patrol
  - patrol ***+VI***
- the
  - the ***+DET***
- temporary
  - temporary ***+ADJ***
- no-fly
  - no-fly ***+ADJ (guessed)***
- zone
  - zone ***+NOUN***
- around
  - around ***+PREP***
- New
  - New ***+PROP***
- Jersey
  - Jersey ***+PROP***
- 's
  - 's ***+POSS***

ZZCad

Helicopters	will	patrol	the	temporary	no-fly	zone	around	New	Jersey	'	s	MetLife	Stadium	Sunday	,	with	F-16s
								proper	proper	'	s	proper	proper	noun	,		proper
noun			det	adj	adj	noun	prep	NP							OCON	prep	NP
SUBJ	v-aux	verb	NP[patrol]				PRP[zone,patrol]										

Current Solutions - Issues

- Common Issues Among Parts of Speech Taggers:
  1. Output
     1. Difficult to Consume
     2. Static
  2. Too Many Parts of Speech Identified
  3. Not Intended for a Non-Technical Audience

NNPS/ Helicopters MD/ will NN/ patrol DT/ the JJ/ temporary JJ/ no-fly NN/ zone IN/ around NNP/ New NNP/ Jersey POS/ 's NNP/ MetLife NNP/ Stadium NNP/ Sunday ,/ , IN/ with NNP/ F-16s VBN/ based IN/ in NNP/ Atlantic NNP/ City JJ/ ready TO/ to VB/ be VBN/ scrambled IN/ if DT/ an JJ/ unauthorized NN/ aircraft VBZ/ does VB/ enter DT/ the VBN/ restricted NN/ airspace ./ .

IN/ Down IN/ below ,/ , JJ/ bomb-sniffing NNS/ dogs MD/ will NN/ patrol DT/ the NNS/ trains CC/ and NNS/ buses WDT/ that VBP/ are VBN/ expected TO/ to VB/ take RB/ approximately CD/ 30,000 IN/ of DT/ the JJ/ 80,000-plus NNS/ spectators TO/ to NNP/ Sunday POS/ 's NNP/ Super NNP/ Bowl IN/ between DT/ the NNP/ Denver NNS/ Broncos CC/ and NNP/ Seattle NNP/ Seahawks ./ .

Solution Statement

ESL COLRS is a computer program that will color different parts of speech to help facilitate ESL students' recognition and comprehension of parts of speech, as relevant to the intent or meaning of a sentence, thus improving their reading efficiency and comprehension by providing a user friendly interface which renders the identification of parts of speech in an easy to consume format.

COLRS Explained

- Utilize Open Source Part of Speech Taggers as our “Engine”
- Output Text in a Clean and Easy to Consume Format
- Allow for User Intervention to Change the P.O.S., if needed
- Save Final Versions of Documents for Later Use

Competition

	Colors	Stanford	Illinois	Children's games (Grammar Gorillas, British Council Games)	ZZCad	Xerox
Ease of use	✓		✓	✓		✓
POS Recognition	✓	✓	✓		✓	✓
POS marking (color)	✓		✓	✓		
Easy to read output	✓		✓	✓	✓	
Editable output	✓					
Target audience (ESL students)	✓					
Simple interface	✓		✓	✓	✓	✓
Custom input	✓	✓	✓		**Limited**	✓
Fast & Efficient	✓	✓	✓		✓	✓

Customer Risks

	Impact				
Probability	1	2	3	4	5
1	C2	C4			C1
2					
3				C3	
4					
5					

C1: Customer adoption
C2: User adoption
C3: Interface complexity
C4: Limited output

Customer Risk Mitigation

	Impact				
Probability	1	2	3	4	5
1	C2	C4			C1
2					
3				C3	
4					
5					

C1: Mentor involvement
C2: User stories/testing
C3: Simple UI
C4: Necessary output with scalability

Technical Risks

	Impact				
Probability	1	2	3	4	5
1					
2			T3		T1
3					T2
4	T4				
5					

T1: Slash integration
T2: POS tagging
T3: Runtime
T4: Limited import options

Technical Risk Mitigation

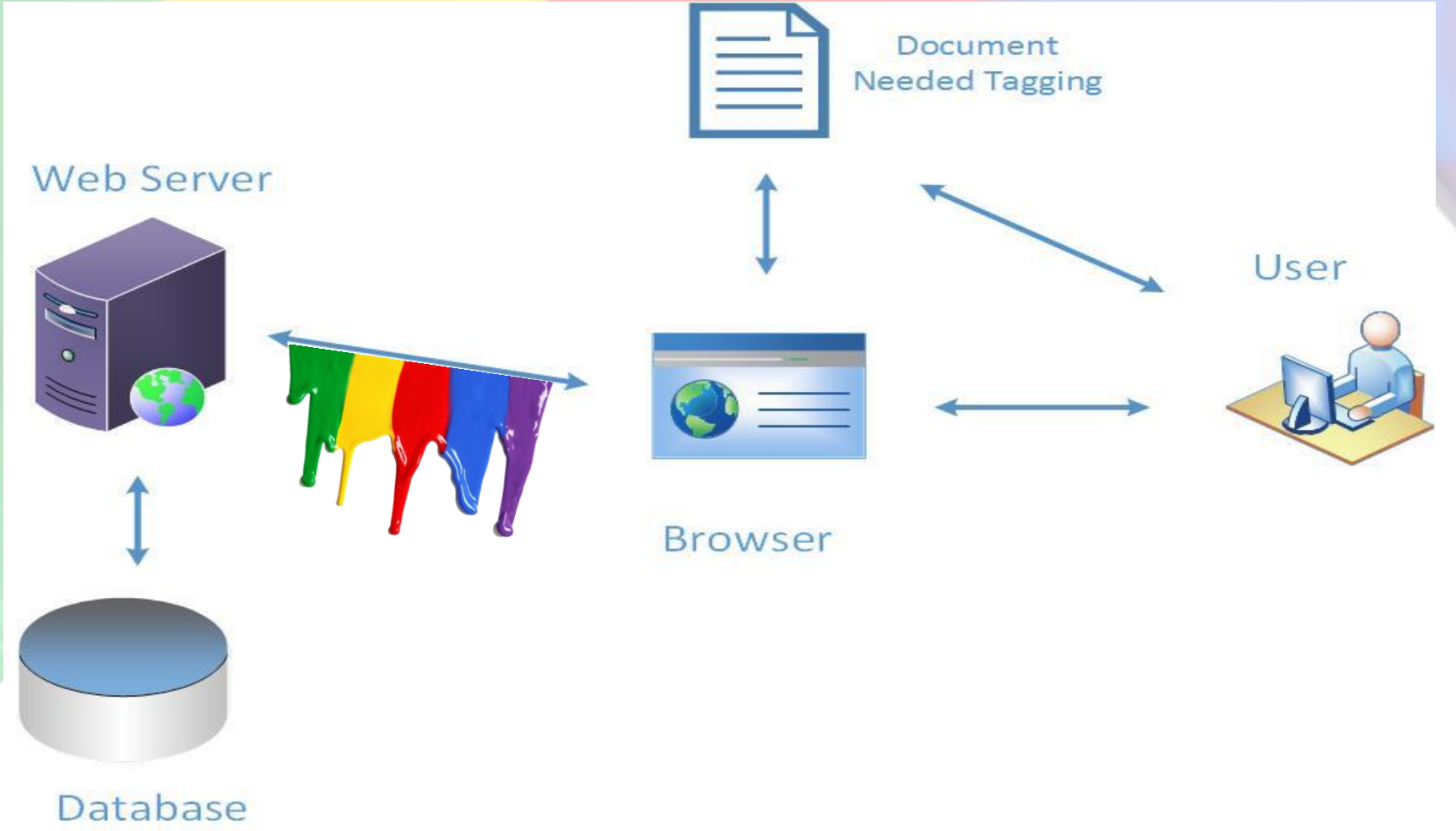
	Impact				
Probability	1	2	3	4	5
1					T2
2			T3		T1
3					
4	T4				
5					

T1: Open communication with ESL Slash
T2: Utilization of existing software
T3: Runtime testing and optimization
T4: Necessary import options with customer approval

What the Objective Entails

- Software capable of identifying parts of speech
- Software ready to accept input and display color coded output
- Software to allow instructors to correct errors from automated parsing
- A data repository of pre-parsed and reviewed documents available for use in lectures or assignments
- “Slash” integration

Initial Architecture



Features Included

Multiple software components integrated to create a simple effective learning interface:

- NLTK (Open source library)
- Node.js
- Skulpt (Open source library embedding python into a web browser)
- Input text processing developed by COLRS team
- Output text processing and UI developed by COLRS team
- Assignment mode with Student Progress Tracking developed by COLRS team

A network based server with backend database, to host all software components, host parsed and reviewed documents, and host the interface for the end users and instructors. Initial development under a local host or intranet.

Wish List

- Beyond plain text or urls of HTML accept additional input formats such as incorporating Office Suites' formats eg: LibreOffice or MS Office, PDFs, RTF, Latex
- Fully web based system with multiple users from divergent locations, including authentication and databases of content for each user and enhanced systems for progress tracking of proficiency improvement.

Batteries Not Included

What will not be included in the project and why.

Your computer is unable to think.

For instance:
Do you love banana pudding?
Your computer hates it.

What Will WE Do?

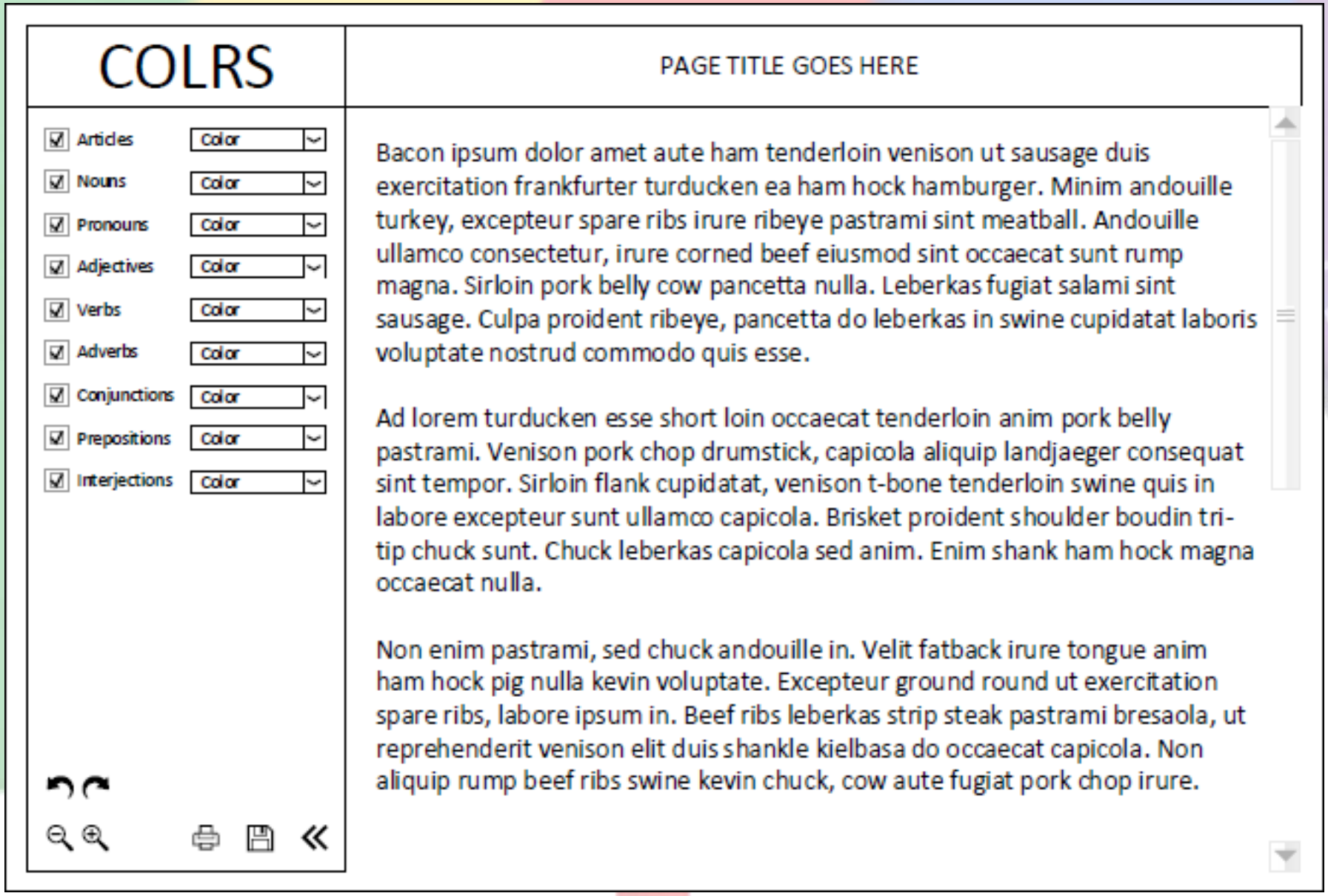
- Design and Build User Interface
  - Including the ability to edit
- Interface with POS Tagging System

Problems

- Accuracy
- Users may have poor grasp of English language

Benefits of COLRS

- Easier to use
- Faster than current solution
- Flexible Output
- Ability to save and edit results
- Track progress
- Reduce redundancy of work by professors


COLRS
PAGE TITLE GOES HERE
Articles
Color
Nouns
Color
Pronouns
Color
Adjectives
Color
Verbs
Color
Adverbs
Color
Conjunctions
Color
Prepositions
Color
Interjections
Color
Bacon ipsum dolor amet aute ham tenderloin venison ut sausage duis exercitation frankfurter turducken ea ham hock hamburger. Minim andouille turkey, excepteur spare ribs irure ribeye pastrami sint meatball. Andouille ullamco consectetur, irure corned beef eiusmod sint occaecat sunt rump magna. Sirloin pork belly cow pancetta nulla. Leberkas fugiat salami sint sausage. Culpa proident ribeye, pancetta do leberkas in swine cupidatat laboris voluptate nostrud commodo quis esse.
Ad lorem turducken esse short loin occaecat tenderloin anim pork belly pastrami. Venison pork chop drumstick, capicola aliquip landjaeger consequat sint tempor. Sirloin flank cupidatat, venison t-bone tenderloin swine quis in labore excepteur sunt ullamco capicola. Brisket proident shoulder boudin tri-tip chuck sunt. Chuck leberkas capicola sed anim. Enim shank ham hock magna occaecat nulla.
Non enim pastrami, sed chuck andouille in. Velit fatback irure tongue anim ham hock pig nulla kevin voluptate. Excepteur ground round ut exercitation spare ribs, labore ipsum in. Beef ribs leberkas strip steak pastrami bresaola, ut reprehenderit venison elit duis shankle kielbasa do occaecat capicola. Non aliquip rump beef ribs swine kevin chuck, cow aute fugiat pork chop irure.

COLRS

PAGE TITLE GOES HERE

Bacon ipsum dolor amet tongue bresaola est short ribs, consequat shank dolore in tri-tip capicola bacon pancetta labore porchetta frankfurter. Pastrami incididunt pancetta, voluptate pork labore bacon. Rump sirloin cupidatat laboris meatloaf nisi pork occaecat. Eu chicken aute, turkey beef in tail nulla.

Deserunt capicola sirloin meatball, dolor tenderloin dolore eu reprehenderit landjaeger beef ribs aute spare ribs. Officia in aliqua kielbasa. Frankfurter ad anim flank fatback. Ut tail officia in fatback, eiusmod voluptate capicola dolore irure adipisicing prosciutto. Nulla non t-bone porchetta adipisicing salami shank veniam. Excepteur tempor exercitation cupidatat veniam sunt tail nisi voluptate capicola venison culpa flank. Tongue meatball t-bone rump reprehenderit turducken shoulder.

Nulla eiusmod qui, brisket chicken strip steak bacon landjaeger chuck. Sausage turkey tenderloin qui ea commodo pork tri-tip nisi salami ut aliqua doner. Hamburger pork chop ullamco dolor mollit. Ground round aliquip commodo, rump flank capicola jowl in meatloaf. Veniam drumstick in swine pancetta lorem pariatur proident nulla andouille. Strip steak deserunt doner ut, veniam boudin et excepteur aute.

Frankfurter elit corned beef turducken bacon tenderloin. Pancetta laboris magna, et pariatur in capicola ribeye kielbasa pig shoulder tenderloin. Sint turducken short loin, commodo chicken hamburger leberkas ea ad et. In dolore leberkas landjaeger commodo frankfurter cillum irure chuck. Tempor duis non elit t-bone sed nulla. Nostrud sausage pariatur ullamco.

Shankle non commodo jowl. Turkey enim brisket sirloin officia. Meatloaf drumstick meatball chuck duis labore. Incididunt prosciutto sed nulla voluptate, reprehenderit enim elit. Dolor leberkas id ut sirloin

Conclusion

References

https://www.odu.edu/esl/courses/intensive

Professor Greg Raver-Lampman – Mentor

http://www.desktopwallpaperhd.net/view/golden-bridge-screen-savers-pictures-francisco-travel-above-california-191995.html

https://www.youtube.com/watch?v=g10jFL423ho

Curitibainenglish.com.br

http://www.wes.org/ras/TrendInInternationalStudentMobility.pdf

http://www.census.gov/compendia/statab/2012/tables/12s0282.pdf

http://grammar.about.com/od/partsofspeech/a/partsofspeech.htm

http://ww2.odu.edu/ao/ira/factbook/cds/data/CDSRDS1314FINAL.pdf

http://japan.usembassy.gov/e/visa/tvisa-niv-fmfaq.html

http://en.Wikipedia.org/wiki/Natural_language_processing

http://en.Wikipedia.org/wiki/Part-of-speech_tagging

http://www.gr8ambitionz.com/2012/05/parts-of-speech-identification.html

Funny-pictures.picphotos.net

References 2

http://www.wes.org/ras/TrendsInInternationalStudentMobility.pdf

http://www.census.gov/compendia/statab/2012/tables/12s0282.pdf

http://grammar.about.com/od/partsofspeech/a/partsofspeech.htm

http://ww2.odu.edu/ao/ira/factbook/cds/data/CDSRDS1314FINAL.pdf

http://japan.usembassy.gov/e/visa/tvisa-niv-fmfaq.html