LEARNER CORPORA OF ENGLISH: GLIMPSES INTO LEARNERS' L2 DEVELOPMENT

Marina Mattheoudakis

Learner corpora

- Collections of electronic texts produced by learners of a foreign language (Hunston 2000).
- What are they used for?
  - to identify patterns of the L2 used by learners,
  - To describe differences between native and non-native linguistic systems,
  - to trace the development of interlanguage,
  - to identify possible L1 interference, and
  - to make comparisons between individual L2 learners or groups of learners.

- HUNSTON Susan, 2000, *Corpora in Applied Linguistics*, Cambridge, Cambridge University Press.

Learner corpora: a bridge?

- A younger field of corpus research
- It is one way of providing a bridge between SLA and FLT
- Learner corpus researchers become progressively aware of the importance of SLA theory and SLA researchers acknowledge the potential value of learner corpora

Major assets of LC

- It brings to the SLA field a much wider empirical basis than has ever previously been available (errors can’t be random)
- LC are more representative of learners’ interlanguage
- The contextualised discourse that learners produce enables researchers to tackle a much wider range of topics
- All language aspects can be studied (Cobb 2003)

Design criteria

- Learner corpora need to be assembled on the basis of very strict design criteria, e.g. timed vs. untimed essay writing, speech vs. writing, learner variables, task variables, etc.

Design criteria for learner corpora

- Language
- Medium
- Text types
- Level of learners
- L2 of learners
- Type of language acquisition (instructed and/or naturalistic)
- Task setting (timed/untimed writing, use of reference tools, etc.)

The design of LC allows us to

- Study the influence of a particular factor (e.g. learners' proficiency level, their L1, the medium, the text type, etc.) on learner language
- With a comparable native speaker corpus, over- and underuse can be studied in addition to mistakes and correct forms

Potential of learner corpora

- Major advantage: they are computerized
- Language “looks rather different when we look at a lot of it at once” (Sinclair 1991:100)
- Real production data rather than experimental data (e.g. grammaticality judgment tasks)
  - drawing conclusions about what a learner can produce spontaneously is difficult on the basis of experimental data

More advantages

- As opposed to experimental data that allow investigations into only a few specific aspects of learner language
  - with learner corpora many aspects can be investigated at once
  - More general questions such as the frequency of different types of mistakes can be addressed
- Aspects of pragmatics and discourse can be studied more easily
- It is not necessary to have a hypothesis prior to the analysis

Limitations

- Learners’ receptive abilities cannot be investigated
- How certain learners are about the acceptability of what they are producing is not known
- If a word or a pattern does not occur, we cannot know whether this is due to ignorance or chance (avoidance strategy, Schachter, 1974)

Learner corpora around the world

- http://www.uclouvain.be/en-cecl-lcworld.html
- The biggest learner corpora to date:
  - The Hong Kong University of Science and Technology (HKUST) Learner Corpus
  - 25 million words and is still growing (Chinese university students learners of English).
  - The Cambridge Learner Corpus; growing all the time
  - over **200,000 exam scripts** from students speaking 148 different languages living in **217 different countries or territories**.

International Corpus of Learner English

- ICLE (International Corpus of Learner English): corpus of electronic texts written by higher intermediate to advanced learners of English of different L1 backgrounds designed by the University of Louvain (Granger et al., 2002).
- Participants represent 17 different mother tongue backgrounds (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Greek, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish and Tswana).

Greek learner corpora

- (1) GRICLE (Hatzitheodorou & Mattheoudakis, 2009)
- (2) Teenage Greek Learner Corpus (Xargia, 2013)
- (3) Experimental Learner Corpus (ELC) (young Greek learner corpus compiled by the Experimental school) (Chasioti, 2013)
- (4) YoLeCorE: Young Learner Corpus of English (audiovisual corpus of instructed learning) (Zapounidis, to appear)

YoLeCorE

- ELT sessions of a whole school year (Oct. 2012-June 2013) were videotaped
- Daily record of whatever is listened to, read, spoken and written by all learners
- MANUAL transcription of the videos with TRANSANA software

Number of tokens per skill

LISTENING	444.312
SPEAKING	581.608
READING	977.516
WRITING	68.288

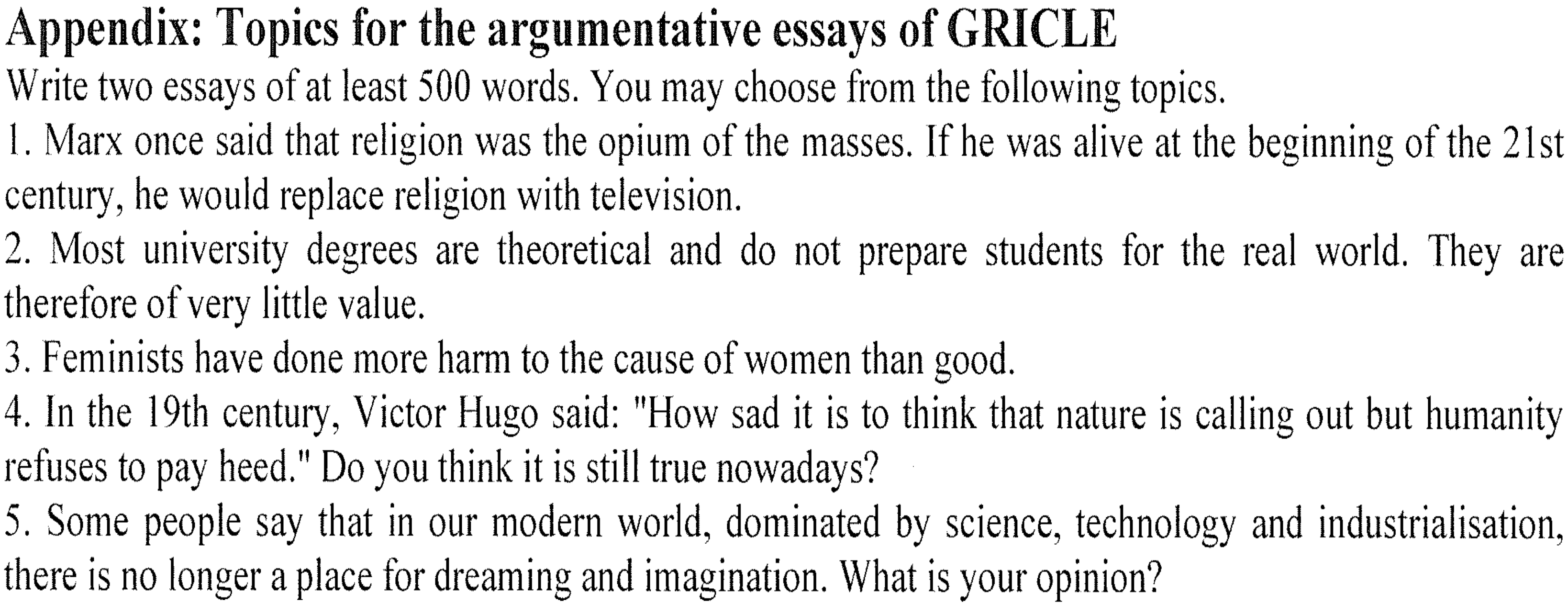
Experimental Learner Corpus

- The ELC consists of 68,832 tokens in total. It comprises 805 different written texts produced during the academic years 2010-2011 and 2011-2012 by a total of 156 students attending the 3 last grades of primary school.

Our learner corpus: GRICLE

- The Greek Corpus of Learner English (GRICLE) : the Greek written component of ICLE; compiled following the guidelines of ICLE
- Participants: 200 Greek native speakers at the 3rd and 4th year of their university studies at the School of English, Aristotle University of Thessaloniki in Greece (age range between 20 and 22 years).
- Each student was required to produce two argumentative essays of at least 500 words each on a given set of topics
- The procedure was timed and students were allowed to have access to reference tools (dictionaries, grammars, etc.).
- The size of GRICLE is approximately 234,000 tokens

Topics in GRICLE



Appendix: Topics for the argumentative essays of GRICLE

Write two essays of at least 500 words. You may choose from the following topics.

1. Marx once said that religion was the opium of the masses. If he was alive at the beginning of the 21st century, he would replace religion with television.

2. Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.

3. Feminists have done more harm to the cause of women than good.

4. In the 19th century, Victor Hugo said: "How sad it is to think that nature is calling out but humanity refuses to pay heed." Do you think it is still true nowadays?

5. Some people say that in our modern world, dominated by science, technology and industrialisation, there is no longer a place for dreaming and imagination. What is your opinion?

Native speaker corpora

- In all studies with GRICLE, two native speaker sub-corpora were used as control of the native writer’s norm
  - LOCNESS (Louvain Corpus of Native English Essays)
  - PELCRA (Polish and English Language Corpora for Research and Applications)

LOCNESS: Louvain Corpus of Native English Essays

- Corpus of native English essays made up of:
  - British pupils’ A level essays: 60,209 words
  - British university students essays: 95,695 words
  - American university students’ essays: 168,400 words
- Total number of words: 324,304 words
- The number and size of essays produced by each student were similar in both GRICLE and LOCNESS; the topics used were selected from the same list

PELCRA project

- PELCRA Corpus was compiled by Leńko-Szymańska (2006)
- American 1st and 2nd-year students; the essays were timed and written in class on a particular topic (size: 25,467 words)

Research into learner language through GRICLE

Contrastive Interlanguage Analysis

- Our studies into learner corpora involve quantitative and qualitative comparisons between native language and learner language (L1 vs L2): Contrastive Interlanguage Analysis
- Studies on
  - how students project their attitude as writers as compared to native speakers
  - Lexical preferences (individual items, lexical chunks, collocations, idioms) by the two groups of writers (native vs non-native)

Research into learner language

- Projection of attitude; construction of coherence (**additive and contrastive adverbial connectors**, e.g., *however, furthermore* & **lexical chunks**, e.g., *it is true that, it is a fact that*)
- Use of **discourse organizing nouns** to construct argumentation (e.g. *argument, statement, opinion*)
- Projection of stance through the use of **stance exponents** (boosters, hedges and attitude markers)
- Lexical patterning of light verbs: *Make, give*, and

Discourse organizing nouns

- How argumentation is constructed through nouns in argumentative essays by apprentice writers, native and non-native speakers of English (Mattheoudakis & Hatzitheodorou, 2010)
- The main function of d.o. nouns: to organise discourse and project stance

Categorization according to function

- They are categorized according to their function into
  - (a) illocutionary, e.g., *statement, argument, point*
  - (b) language activity, e.g., *dispute, debate, controversy*
  - (c) mental process, e.g., *opinion, view* (Francis 1994)

Results

illocutionary	Native corpora	GRICLE
Argument	253	19
Statement	49	43
point	46	19

Language activity nouns

Language activity	Native corpora	GRICLE
Debate	57	8
Controversy	10	0
dispute	2	0

Mental process nouns

Mental process nouns	Native corpora	GRICLE
Idea	72	33
View	37	26
opinion	16	27

Overall differences

Nouns	Native Corpora	GRICLE
Illocutionary	348	81
Language activity	150	91
Mental process	58	2

It seems that

- Greek learners prefer to use adverbials (e.g., *however, in addition)* in order to express connectivity
  - Teacher induced????
- Teaching of lexical cohesion for the production of argumentation is largely neglected; Discourse organizing nouns are rarely discussed
- Perhaps the use of d.o. nouns to refer backwards and forwards to a proposition and label requires increased mental processing

The impact of culture on the use of stance exponents (Hatzitheodorou & Mattheoudakis, 2009, 2011)

- This study looked into the projection of stance through the use of **stance exponents** (boosters, hedges and attitude markers) in GRICLE and native corpora
- It also examined the extent to which learners' written production is influenced by cultural factors

Categories of stance features

- **Hedges**: *possibly, probably, may, might, maybe, presumably, relatively,* etc.
- **Boosters**: *it is evident that, it is clear that, it is a fact that, it is true that, it is obvious that, clearly, evidently, obviously, definitely, certainly, truly,* etc.
- **Attitude markers – effect**: *I feel, I hope, it amazes me, it surprises me, it is shocking, it is (un)fortunate, (un)fortunately, happily, luckily,* etc.
- **Attitude markers – opinion**: *I think, I agree, I believe, I consider, I gather, I conclude*, etc.
- **Self-mention**: *I, we my, our*
  - *Adapted model of stance by Hatzitheodorou & Mattheoudakis, 2011*

Results (1)

- Boosters in GRICLE get the lion’s share as their use is much more extensive than that of hedges and attitude markers (**334 occurrences of boosters, 112 of hedges, 95 of attitude markers**).
- Moreover, boosters are much more frequent in GRICLE than in the native corpora (**334 vs 163** occurrences respectively).

Boosters

Boosters	GRICLE	Native Corpora
Of course	N=153; 8.62/10,000	N=34; 1.94/10,000
No doubt/undoubtedly without any doubt	N=63; 3.54	N=13; 0.74
Indeed	N=46; 2.59	N=15; 0.86
Definitely	N=12; 0.66	N=28; 1.59
Truly	N=14; 0.78	N=24; 1.37
Clearly	N=9; 0.5	N=27; 1.54

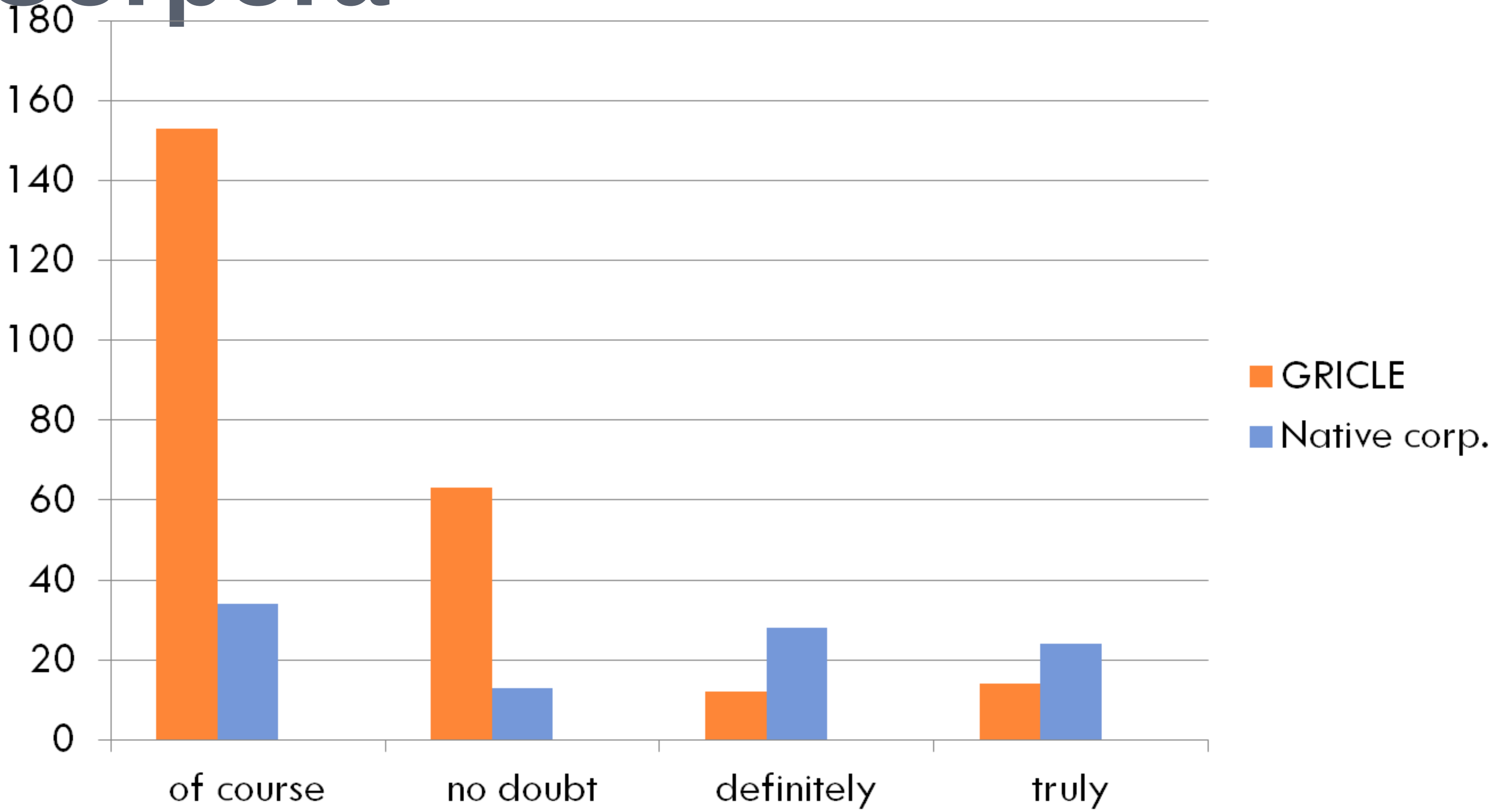
Boosters

- Greek learners tend to use lexical chunks as boosters much more frequently than NS (**160 vs 24**): *it is true that, it is a fact that, it is obvious that*

Ex: *It is clear and obvious that today's society would be different from its foundations if television had not existed.*

- With respect to adverbs used as boosters, differences between Greeks and native speakers are less striking; there is variability depending on the adverb

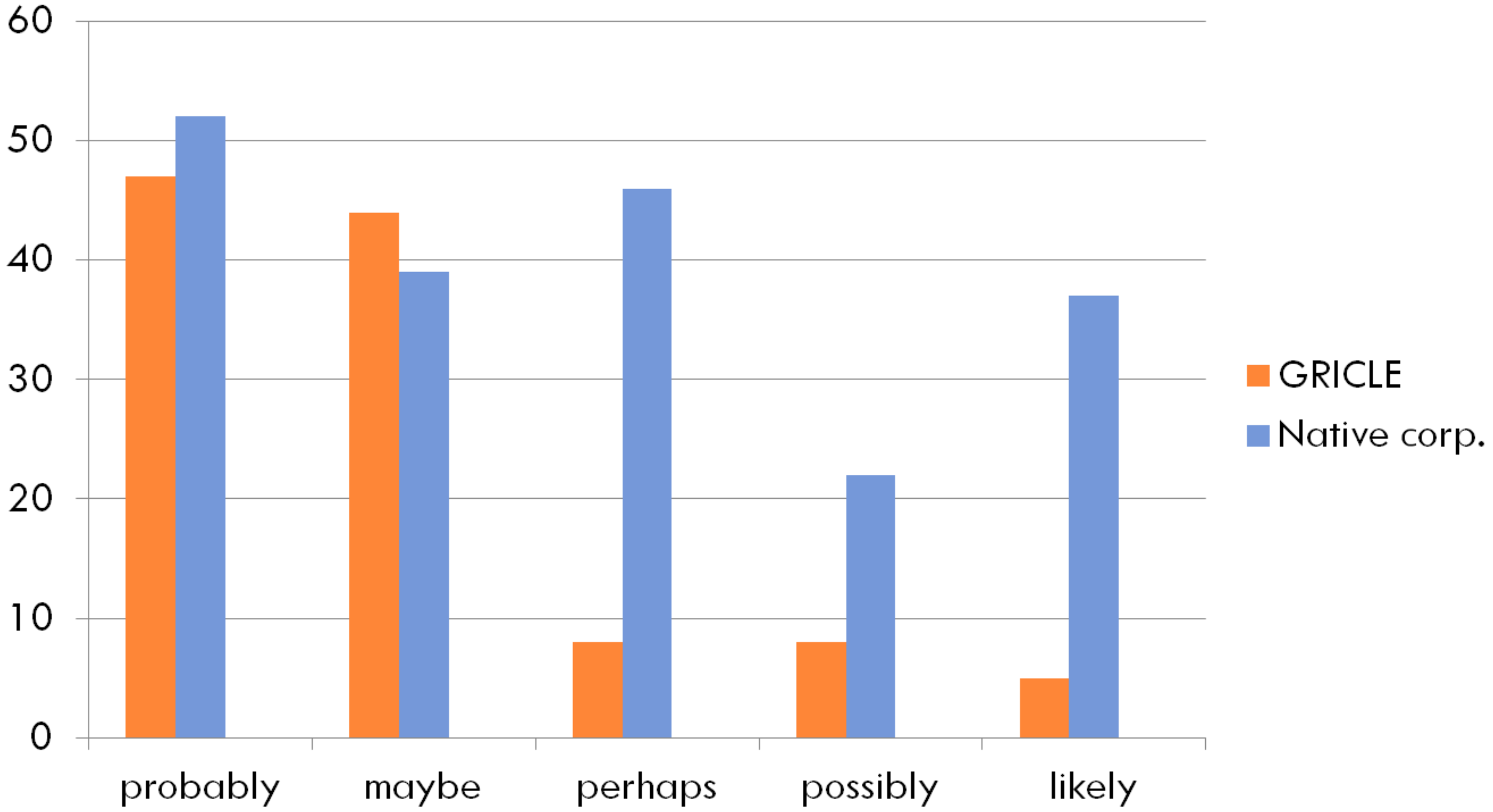
Adverbs as boosters in GRICLE and in Native Corpora



Results (2)

Hedges	GRICLE	Native Corpora
Probably	N=47; 2.64/10,000	N=52; 2.97/10,000
Maybe	N=44; 2.48	N=39; 2.22
Perhaps	N=8; 0.45	N=46; 2.63
Possibly	N=8; 0.45	N=22; 1.26
Likely	N=5; 0.28	N=37 2.11

Hedges in GRICLE and in Native Corpora



Hedges

- Hedging features more often in the native corpora than GRICLE.
- Anglo-American rhetorical convention: overstatements are generally avoided (cf. Hyland, 2005)

Examples

- *Our society and the feminists that support this equality is so intent on creating this type of environment, that perhaps we have become obsessed with just that and ultimately losing a part of what should be a unique creation. We were created different and we are different. Maybe we are different for a reason and those differences should not necessarily be viewed as negative. Maybe we were made different to stay different and perhaps in trying to create this equality we may lose something very unique and special that can never be regained. What consequences might there be for us that we cannot undo? (LOCNESS) [Appendix II, topic 5].*

Cultural factors

- Compared to native speakers, Greeks wish to project a confident attitude and believe that, to achieve that, they need to make frequent use of boosters (undoubtedly, of course, etc.); hence, their more emphatic writing.
- According to Hofstede (1980), Greeks do not favour uncertainty and this is reflected in their tendency to be emphatic and avoid hedging in writing.

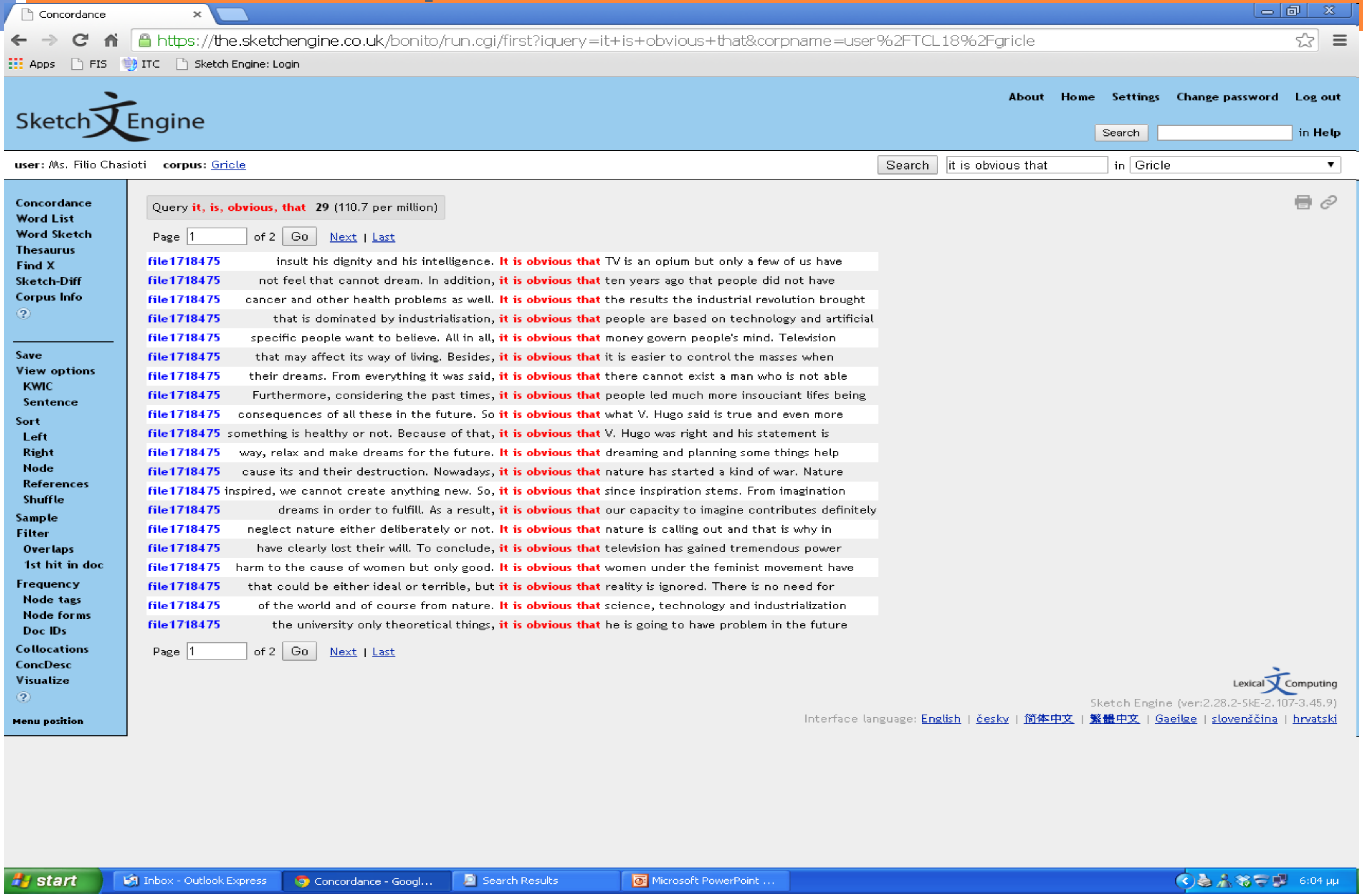
So,

- Greek learners’ tendencies may be attributed to a combination of factors related to
  - (a) L1 and L2 instruction materials and techniques during the learners’ secondary and tertiary education,
  - (b) transfer of learners’ L1 style of writing and cultural features,
  - (c) their development as L2 writers.

A look at the corpus through sketchengine

- Online corpus query tool, www.sketchengine.co.uk, developed by Adam Kilgarriff, Pavel Smrz and David Tugwell
- “[t]he software was originally designed as a tool for dictionary makers,
- Sketch Engine allows access to a number of corpora and users are able to upload their own raw data and compile a corpus themselves (Pearse, 2008, p. 4; Kilgarriff et al., 2004).

Comparative data: GRICLE vs native corpora



https://the.sketchengine.co.uk/bonito/run.cgi/first?iquery=it+is+obvious+that&corpname=user%2FTCL18%2Fgricle

user: Ms. Filio Chasioti **corpus:** Gricle

Search it is obvious that in Gricle

Query **it, is, obvious, that** **29** (110.7 per million)

Page 1 of 2 Go Next | Last

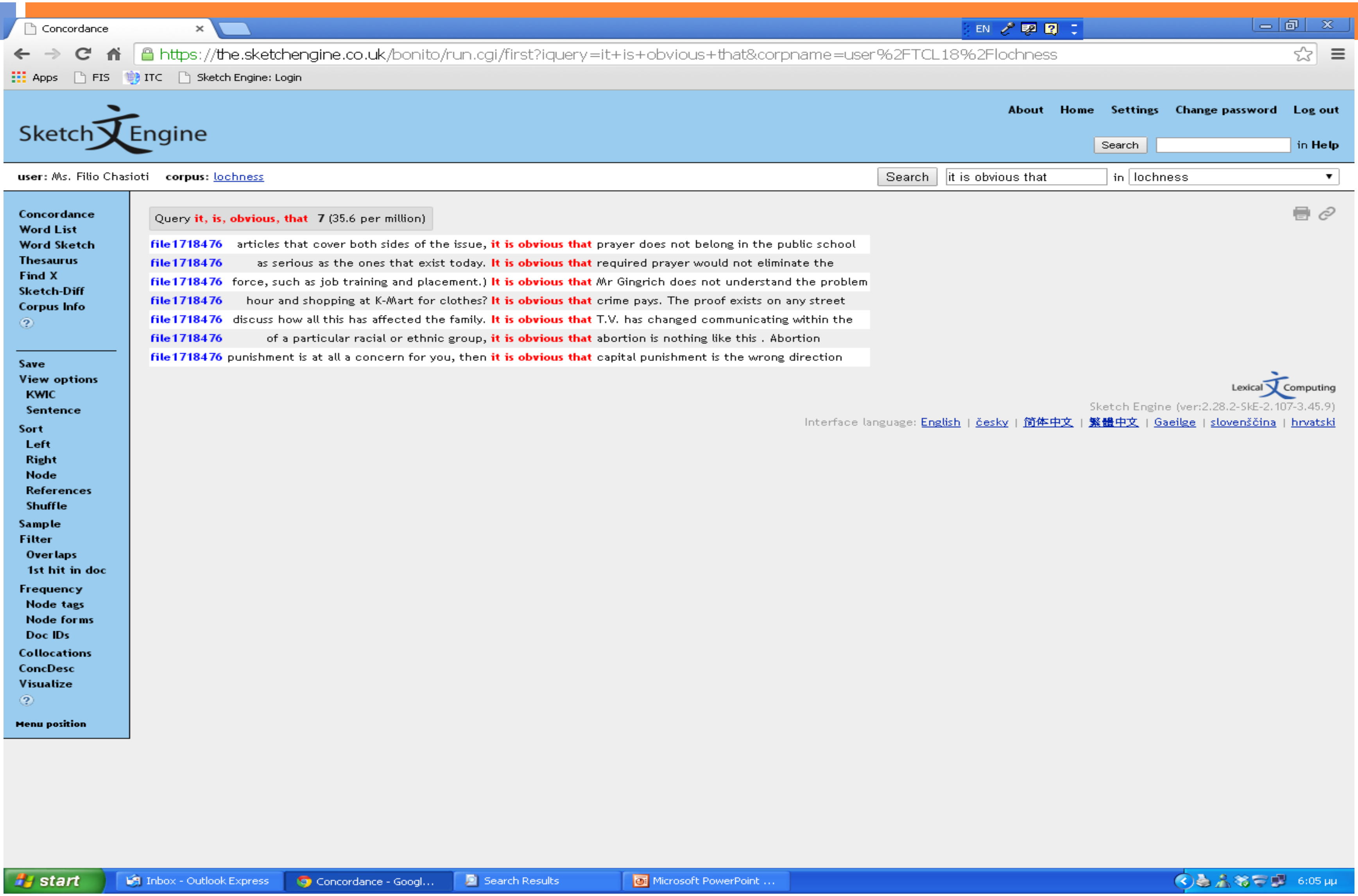
file1718475	insult his dignity and his intelligence.	**It is obvious that**	TV is an opium but only a few of us have
file1718475	not feel that cannot dream. In addition,	**it is obvious that**	ten years ago that people did not have
file1718475	cancer and other health problems as well.	**It is obvious that**	the results the industrial revolution brought
file1718475	that is dominated by industrialisation,	**it is obvious that**	people are based on technology and artificial
file1718475	specific people want to believe. All in all,	**it is obvious that**	money govern people's mind. Television
file1718475	that may affect its way of living. Besides,	**it is obvious that**	it is easier to control the masses when
file1718475	their dreams. From everything it was said,	**it is obvious that**	there cannot exist a man who is not able
file1718475	Furthermore, considering the past times,	**it is obvious that**	people led much more insouciant lifes being
file1718475	consequences of all these in the future. So	**it is obvious that**	what V. Hugo said is true and even more
file1718475	something is healthy or not. Because of that,	**it is obvious that**	V. Hugo was right and his statement is
file1718475	way, relax and make dreams for the future.	**It is obvious that**	dreaming and planning some things help
file1718475	cause its and their destruction. Nowadays,	**it is obvious that**	nature has started a kind of war. Nature
file1718475	inspired, we cannot create anything new. So,	**it is obvious that**	since inspiration stems. From imagination
file1718475	dreams in order to fulfill. As a result,	**it is obvious that**	our capacity to imagine contributes definitely
file1718475	neglect nature either deliberately or not.	**It is obvious that**	nature is calling out and that is why in
file1718475	have clearly lost their will. To conclude,	**it is obvious that**	television has gained tremendous power
file1718475	harm to the cause of women but only good.	**It is obvious that**	women under the feminist movement have
file1718475	that could be either ideal or terrible, but	**it is obvious that**	reality is ignored. There is no need for
file1718475	of the world and of course from nature.	**It is obvious that**	science, technology and industrialization
file1718475	the university only theoretical things,	**it is obvious that**	he is going to have problem in the future

Page 1 of 2 Go Next | Last

Sketch Engine (ver:2.28.2-SkE-2.107-3.45.9)

Interface language: English | česky | 简体中文 | 繁體中文 | Gaeilge | slovenščina | hrvatski

Native corpora



user: Ms. Filio Chasioti corpus: lochness

it is obvious that in lochness

Query **it, is, obvious, that** 7 (35.6 per million)

file1718476 articles that cover both sides of the issue, **it is obvious that** prayer does not belong in the public school

file1718476 as serious as the ones that exist today. **It is obvious that** required prayer would not eliminate the

file1718476 force, such as job training and placement.) **It is obvious that** Mr Gingrich does not understand the problem

file1718476 hour and shopping at K-Mart for clothes? **It is obvious that** crime pays. The proof exists on any street

file1718476 discuss how all this has affected the family. **It is obvious that** T.V. has changed communicating within the

file1718476 of a particular racial or ethnic group, **it is obvious that** abortion is nothing like this . Abortion

file1718476 punishment is at all a concern for you, then **it is obvious that** capital punishment is the wrong direction

Sketch Engine (ver:2.28.2-SkE-2.107-3.45.9)

Interface language: English | česky | 简体中文 | 繁體中文 | Gaeilge | slovenščina | hrvatski

GRICLE

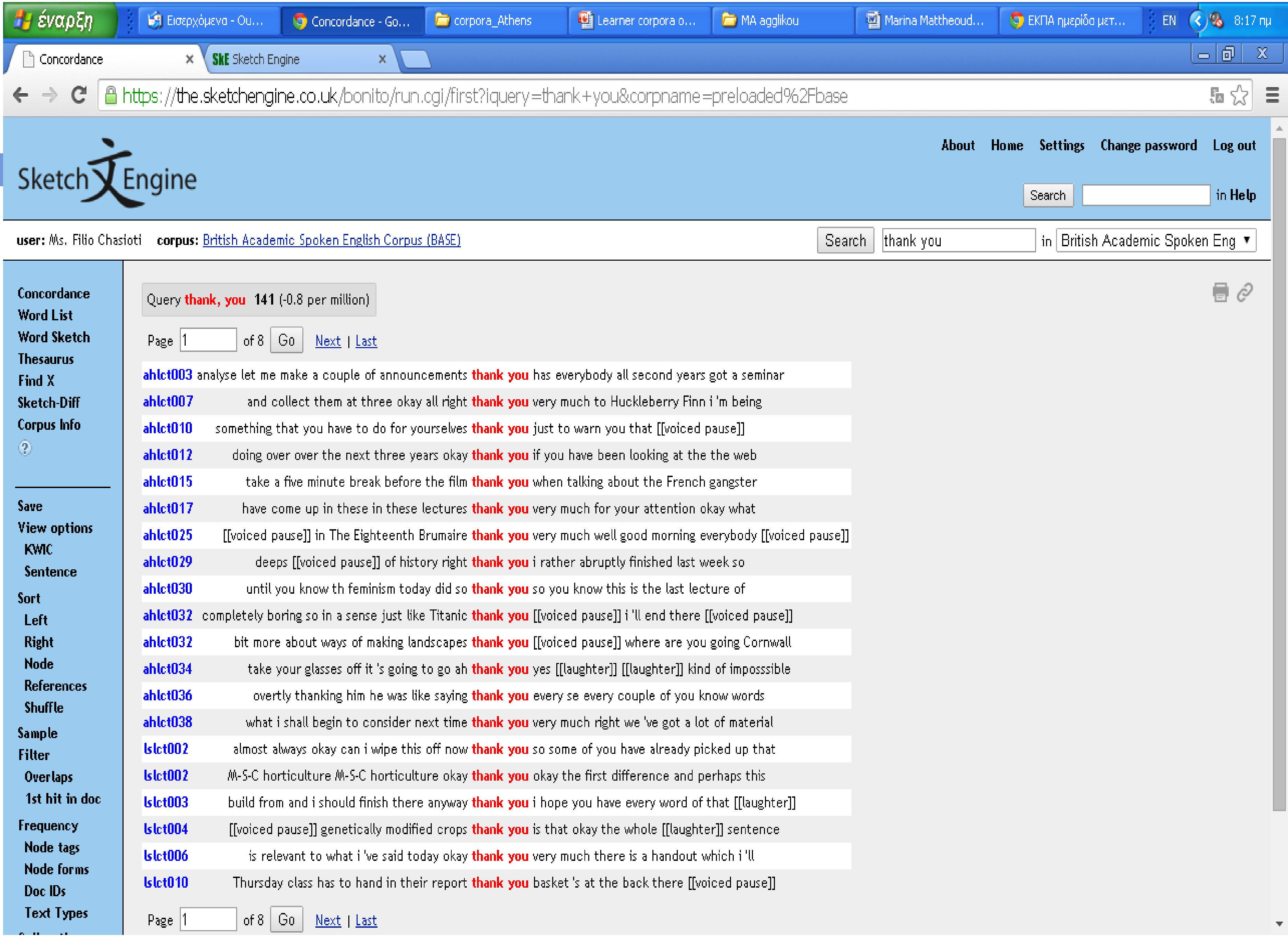
- And due to the blind faith their audience has television and its people have become omnipotent, the fourth force of society as some say; it has the power to judge, question and condemn somebody, or even glorify and create new role-models. Certainly it can make the masses focus on certain issues and bypass others, that can be even more important than the former. The masses have clearly lost their will. To conclude, ***it is obvious that*** television has gained tremendous power over the past 60 years and has evolved in so great a force that has replaced in a way the power that religion could once impose to people.

LOCNESS

- Proponents of prayer in public schools believe that a religious infusion is needed to balance the lack of values and the increasing rate of violence in society. The opponents hold, however, that prayer in public schools would destroy the separation of church and state, and that prayer will not be able to end the ills of society. With the widespread views this debate creates, many writes have taken it upon themselves to offer their opinions on the subject. After considering these articles that cover both sides of the issue, ***it is obvious that*** prayer does not belong in the public school classroom, as the articles that oppose prayer in public schools refute and weaken considerably the arguments for the reintroduction of prayer in public schools as a way to cure modern socia

Learner corpus bibliography

- http://www.uclouvain.be/en-cecl-lcBiblio.html
- For further information regarding Corpus Querying and Grammar Writing for the Sketch Engine, see: http://trac.sketchengine.co.uk/wiki/SkE/CorpusQuerying#



user: Ms. Filio Chasioti corpus: British Academic Spoken English Corpus (BASE)

Query thank, you 141 (-0.8 per million)

Page 1 of 8 Go Next | Last

ahlct003	analyse let me make a couple of announcements	thank you	has everybody all second years got a seminar
ahlct007	and collect them at three okay all right	thank you	very much to Huckleberry Finn i 'm being
ahlct010	something that you have to do for yourselves	thank you	just to warn you that [[voiced pause]]
ahlct012	doing over over the next three years okay	thank you	if you have been looking at the the web
ahlct015	take a five minute break before the film	thank you	when talking about the French gangster
ahlct017	have come up in these in these lectures	thank you	very much for your attention okay what
ahlct025	[[voiced pause]] in The Eighteenth Brumaire	thank you	very much well good morning everybody [[voiced pause]]
ahlct029	deeps [[voiced pause]] of history right	thank you	i rather abruptly finished last week so
ahlct030	until you know th feminism today did so	thank you	so you know this is the last lecture of
ahlct032	completely boring so in a sense just like Titanic	thank you	[[voiced pause]] i 'll end there [[voiced pause]]
ahlct032	bit more about ways of making landscapes	thank you	[[voiced pause]] where are you going Cornwall
ahlct034	take your glasses off it 's going to go ah	thank you	yes [[laughter]] [[laughter]] kind of impossible
ahlct036	overtly thanking him he was like saying	thank you	every se every couple of you know words
ahlct038	what i shall begin to consider next time	thank you	very much right we 've got a lot of material
lslct002	almost always okay can i wipe this off now	thank you	so some of you have already picked up that
lslct002	M-S-C horticulture M-S-C horticulture okay	thank you	okay the first difference and perhaps this
lslct003	build from and i should finish there anyway	thank you	i hope you have every word of that [[laughter]]
lslct004	[[voiced pause]] genetically modified crops	thank you	is that okay the whole [[laughter]] sentence
lslct006	is relevant to what i 've said today okay	thank you	very much there is a handout which i 'll
lslct010	Thursday class has to hand in their report	thank you	basket 's at the back there [[voiced pause]]

Page 1 of 8 Go Next | Last