



# A Corpus Based Investigation into Science Communication

By Newcastle University iGEM Team





# Contents

Glossary of terms.....	3
1. Introduction.....	4
2. Method: Using Corpus Linguistics.....	5
➤ 2.1.COCA.....	5
➤ 2.2 NOW.....	5
➤ 2.3 Advantages of Corpora Used.....	6
➤ 2.4. Tools and Analysis Techniques.....	6
3. Results and Discussion.....	7
➤ 3.1. Genetic Engineering vs Synthetic Biology.....	7
➤ 3.2. GMO.....	12
➤ 3.3. Biosensor.....	14
4. Conclusion.....	15
References and Acknowledgments.....	17

# Glossary of Terms

**COCA: Corpus of Contemporary American English:** a corpus of English containing in excess of 520 million words of text. The corpus features spoken, fiction, magazines and academic texts, from 1990-2015.

**NOW: Corpus of News on the Web:** a corpus featuring texts from web based newspapers and magazines, with 5.1 billion words of data from 2010- present day.

**Frequency:** the number of times a word occurs per million words in a corpus.

**Collocates:** frequently co-occurring words.

**Concordance/ KWIC Lines:** Key Word in Context Lines, which show the result of a search in a corpus in the sentence in which it was used in the source text.





# A Corpus Based Investigation into Science Communication

## 1. Introduction

For advancements in science to have an impact, it is important that they are communicated in a way that the largest possible audience can access and engage with. By looking at how, and where, particular areas of science are communicated, it is possible to gain an awareness of attitudes towards topics, and how vastly they are covered. In turn, this information can be used to determine key issues and areas for improvement, in terms of how the public engage with science through language. In this report, the linguistic techniques of corpus linguistics and discourse analysis will be used to investigate science communication. The purpose of this report is to establish how language reflects attitudes towards scientific topics, with a particular focus on synthetic biology, and how language can be used most effectively in communicating a scientific project. This information feeds back into the Newcastle iGEM project, as it will be used to help produce a set of guidelines for communicating synthetic biology.

Corpus linguistics is a technique used to search large collections of text for patterns, allowing more widespread conclusions to be drawn. As Flowerdew (2013: 160) defines it, corpus linguistics is 'the application of computational tools to the analysis of corpora, in order to reveal language patterns which systematically occur in them'. For this research, corpus linguistics will be used to search four key terms, in two main corpora. These terms are *Genetic Engineering*, *Synthetic Biology*, *GMO*, and *biosensor*. These searches have been chosen because of their relevance to current advancements and the principles of iGEM. *Biosensor* is a search more specific to Newcastle University's iGEM project: creating Sensynova, a biosensor development platform.

An aim of this research is to determine how the public engage with synthetic biology and related topic areas by identifying any patterns in language use in texts that address these issues, and investigating how they are discussed in the media. Previously in iGEM, teams have used surveys to garner a public opinion on GMOs and synthetic biology. Corpus linguistics is an alternative method that has here been used to gain an understanding of public opinion, by understanding how language use reflects attitudes towards a subject.

Alongside corpus linguistics, which has been used to find patterns in language use, discourse analysis will be used to interpret these patterns. Discourse analysis is a method of analysing texts that considers language as a 'social practise'. The focus is on how we use language in order to interact and communicate, adapting to different contexts, rather than language being exclusively a set of rules to adhere to in the transfer of information (Wood and Kroger 2000: 4). In the report, it will be shown how



combining corpus linguistics and discourse analysis is an effective method to assess the social impact of the language used in science communication, as the former technique identifies patterns, while the latter explores reasons for these patterns.

This report will take the following structure. In section 2, the method used, including the corpora and techniques of analysis, will be explained. The following section 3 will feature the results and discussion of the research into each of the key terms: *Genetic Engineering* and *Synthetic Biology*, *GMO*, and *biosensor*. Finally, section 4 finishes with the key conclusions and points for future research, which will feed into a set of guidelines for science communication.

## 2. Method: Using Corpus Linguistics

For this research, two main corpora from the Brigham Young University corpora collection have been used. These are COCA: The Corpus of Contemporary American English (Davies, 2008), and NOW: Corpus of News on the Web (Davies, 2013). These corpora have been selected because their features are beneficial to help achieve the aims of this investigation.

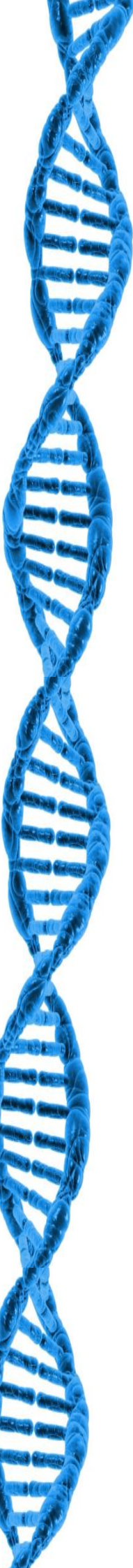
### 2.1 COCA

Firstly, COCA is a large corpus of English and American English, comprising of texts from a number of genres: spoken, academic, fiction, newspaper, and magazines. Therefore, COCA can be used to compare how the search terms feature in different genres, and suggest the significance of any variation. For the purposes of this research, the academic, newspaper, and magazine texts in COCA will be focused on. COCA is very widely used, and contains more than 520 million words of text, from 1990-2015.

### 2.2 NOW

Secondly, NOW contains in excess of 5.1 billion words of data, collected from web-based newspapers and magazines. NOW features texts from 2010- present day; it is an extremely up to date corpus, as it grows daily. Therefore, using NOW gives evidence of how the media currently reports on the key search terms, and hence public opinions and attitudes associated with them. Sampling bias must be considered, and to account for this the whole breath of news sources that NOW collects from have been used in analysis. Therefore, worldwide sources are represented.





## 2.3 Advantages of Corpora Used

COCA and NOW have been used together in this research because the limits of one can be made up by the advantages of the other. For example, because NOW only starts in 2010, it is advantageous to also use COCA, whose records begin in 1990. Any development over time prior to 2010 can then be viewed. On the other hand, as COCA currently only holds records up until 2015, the use of NOW, which is up to date to the writing of this report, allows development after 2015 to be considered. Using the most up to date information available is particularly important when an aim of this research is to gain an understanding of current public opinion.

## 2.4 Corpus Linguistics Tools and Analysis Techniques

Some key corpus linguistics tools have been used in this research. Primarily, frequency data has been retrieved using COCA, showing the number of times each search term occurs in a corpus. This data is then sorted into year categories and genre categories. Frequency is measured in words per million: the frequency equates to the number of times each search term occurs per million words in COCA. This allows comparison across different text sizes. Frequency counts have been represented in graphs in section 3 of this report.

Next, the collocates of each word have been identified using the collocates search function in the NOW corpus. Collocates are words which frequently occur around each other (Evison 2010: 130). The top twenty adjectives which collocate with each search term have been identified: these words occur near the search terms most frequently, and are therefore most significant. The collocate ranked 1 is the most frequent. By setting the search parameters to adjectives that occur up to four places either before or after the search term, it has been ensured that the most relevant words to fulfil the aims of this study are identified- use of modifiers is a strong indicator of attitudes towards a subject. Some manual filtering has also been used. This is to ensure collocates which occur because of limitation of the corpora, and are therefore anomalous, are excluded. For example, a word may appear as a collocate because the same source has been recorded in a corpus multiple times. This will increase the frequency count of the collocate, when really it is the same source which is contributing multiple times.

Finally, concordance lines have been used. When the collocates of the search terms are looked at in isolation, it can be difficult to interpret anything meaningful from them. However, by using the concordance, or key word in context lines (KWIC lines), that the corpus provides, further analysis using principles from discourse analysis can be carried out. KWIC lines provide more context for how the words have been used, to allow the intended impact of the words to be seen. When the collocates are

identified through the search in the corpus, they are taken from the sentence or sentences they were used in in the corresponding source articles. The KWIC lines provide the sentence or sentences that illustrate how the collocate occurs in relation to the original search term.

By combining these techniques, an insight into language use around each of the four categories has been gained. In turn, the language use reflects perceptions and attitudes associated with each topic, as will be discussed in the following sections.

### 3. Results and Discussion

#### 3.1 Genetic Engineering vs Synthetic Biology

The following section features discussion of findings after searching *genetic engineering* and *synthetic biology* in the corpora. It has been useful to group these two search terms together for comparison and discussion. This allows genetic engineering, a more established field than synthetic biology, to be used as a kind of reference point for synthetic biology. Therefore, the impact of synthetic biology in relation to genetic engineering can be viewed, along with where coverage can be improved.

##### 3.1.1 Frequency Over Time, Using COCA

Figure 1 displays how the frequencies of *genetic engineering* and *synthetic biology* compare across time, from the year 1990 to 2015. To account for the anomaly in that the coverage of genetic engineering drops significantly in the years of 2005-2009, compared to 2000-2004, a further search of *genetic modification* has been completed. The feature of *genetic modification* suggests that this is an alternative term which can be used in publications which discuss genetic engineering.

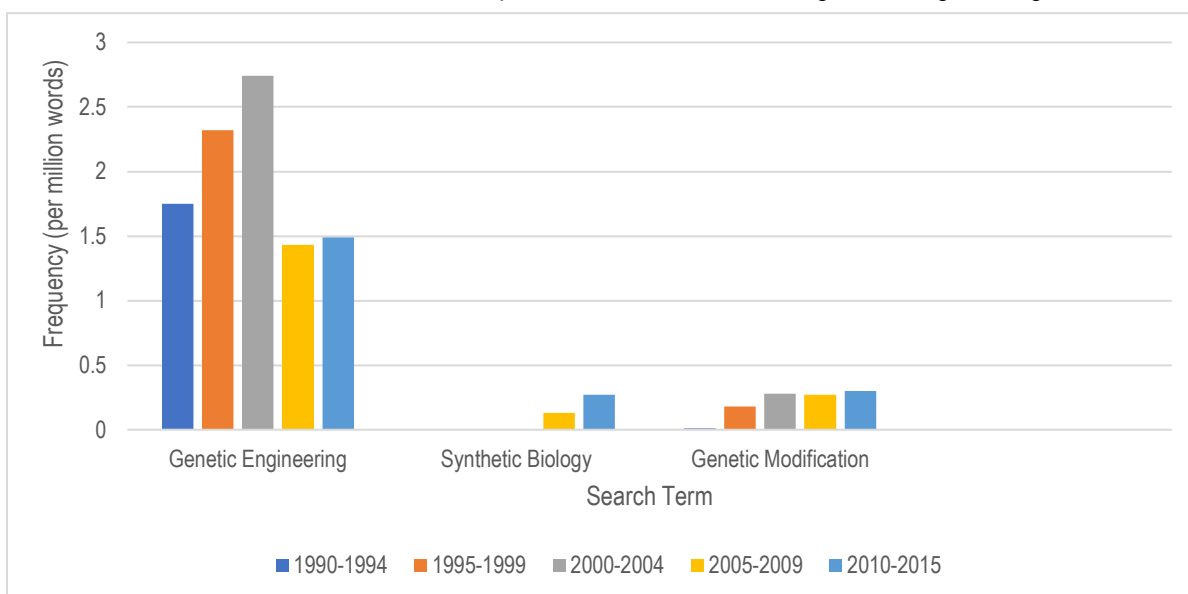


Figure 1: Frequency of *Genetic Engineering*, *Synthetic Biology* and *Genetic Modification* in texts in 2015 (Davies 2008).

As illustrated in Figure 1, Genetic Engineering has been talked about in texts featuring in COCA as far back as its records go- starting in 1990. It increases in frequency for each four-year division, up until the end of 2004, showing the natural progression that as the field of genetic engineering advanced, more people discussed it, and it featured more in texts. After 2004, the frequency of *genetic engineering* drops quite significantly, and starts to level off up to the last of the records that COCA holds, in 2015. It is interesting to note that in the years that the frequency of *genetic engineering* decreases, the frequency of *synthetic biology* rises. Although the rise in the frequency of *synthetic biology* is obviously due to its emergence as a field, this correlation could suggest the start of a shift in focus from an older discipline to a newer one. Despite this, the frequency of *synthetic biology* in the years 2005-2015 is still significantly lower than that of *genetic engineering*, indicating that coverage of synthetic biology is low, and consequently, this is a sign that it does not reach as large an audience. Thus, the results indicate that for synthetic biology to reach a larger audience, its frequency must continue to increase. As will be discussed in the following sections, reaching a larger audience can have both a positive and negative impact, based on the ways the field is reported on.

### 3.1.2 Frequency over genres in COCA

Figure 2 shows the frequency of *genetic engineering* and *synthetic biology* across the genres of magazine, newspaper, and academic text, in COCA.

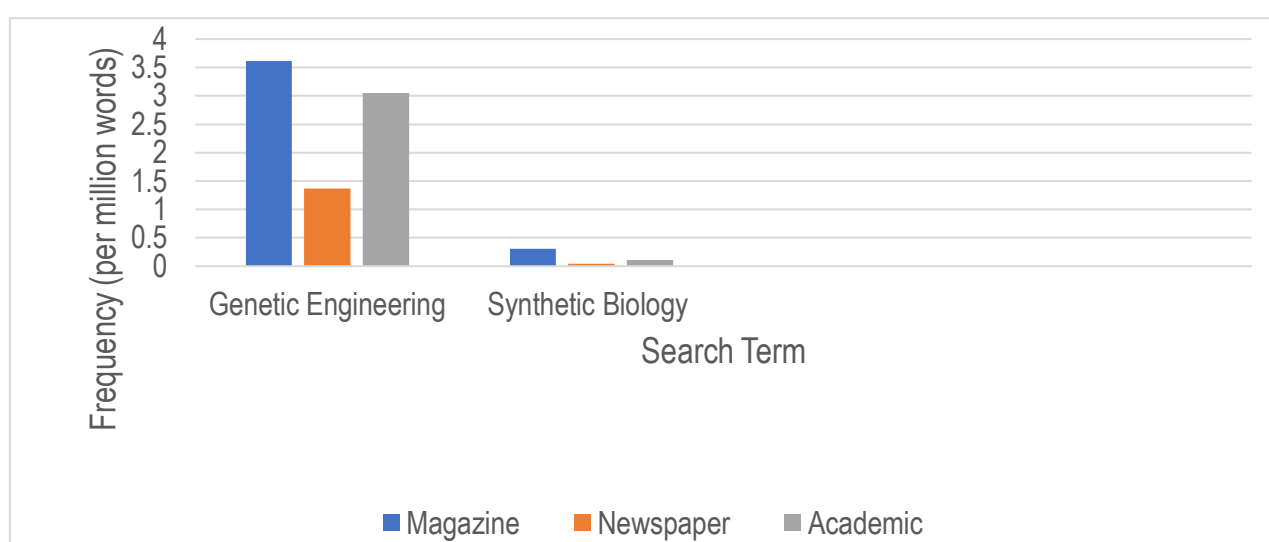


Figure 2: Frequency of *Genetic Engineering* and *Synthetic Biology* across the genres of magazine, newspaper, and academic publications in COCA (Davies 2008).

One of the most interesting things to note is that both *genetic engineering* and *synthetic biology* feature more frequently in magazines compared to newspapers. By using the feature of COCA which provides the name of the publication which the search term is featuring in, it becomes clear that the higher



frequency in magazines reflects the contribution of field specific magazines, such as *Popular Science*, *Science News* and *Horticulture*. The discussion around genetic engineering and synthetic biology features predominantly in publications that people who are already invested and interested in these areas of science are more likely to read. For these disciplines to be reaching a wider audience, it would be hoped that the frequency of the terms in newspapers will increase in future years, as this is a genre with less specific feature focus. The low frequency of *synthetic biology* in academic articles, compared to that of *genetic engineering* further emphasises how this is an emerging field.

### 3.1.3 Analysis of *Genetic Engineering* and *Synthetic Biology* collocates

Figures 3 and 4 display the top 20 collocates in the NOW corpus, with some manual filtering to discount anomalies, for the search terms *genetic engineering* and *synthetic biology*, respectively.


Ranking of Collocate	Collocate
1	International
2	New
3	Human
4	Natural
5	Molecular
6	Artificial
7	Modern
8	Other
9	Central
10	Conventional
11	Synthetic
12	Different
13	Global
14	Agricultural
15	Extreme
16	Traditional
17	Nuclear
18	Possible
19	Advanced
20	Controversial

Figure 3: Top 20 adjectives which collocate with *Genetic Engineering*, in NOW corpus (Davies 2013).

Ranking of Collocate	Collocate
1	New
2	Biological
3	Emerging
4	Genetic
5	Global
6	Synthetic
7	Artificial
8	Evolutionary
9	Industrial
10	Metabolic
11	Potential
12	Brave
13	Future
14	Other
15	Clinical
16	International
17	Natural
18	Cutting-Edge
19	Advanced
20	Ethical

Figure 4: Top 20 adjectives which collocate with *Synthetic Biology*, in NOW corpus (Davies 2013).

When considered on their own, it is harder to see the full significance of the collocates. However, by using the KWIC lines, each collocate can be viewed in the context of the text it features in, and principles from discourse analysis can be used to draw conclusions. For example, when looked at in



isolation, some of the adjectives may seem neutral. However, using the KWIC lines shows they are being used with positive or negative evaluative qualities.

Considering the semantic prosody of collocates gives an indication of whether these adjectives, which may seem neutral, are being used to portray genetic engineering and synthetic biology in a positive or negative light. Louw (1993, cited in Ahmadian *et al*, 2011: 288) defines semantic prosody as ‘a form of meaning which is established through the proximity of consistent series of collocates often characterisable as positive or negative and whose primary function is the expression of the attitude of its speaker or writer toward some pragmatic situation’. This means a word can take on particular associations based on the type of words it frequently occurs with.

Using all of this information, the following conclusions related to the portrayal of synthetic biology and genetic engineering have been drawn, making comparisons where appropriate.

The collocates *international* and *global*, which feature for both *genetic engineering* and *synthetic biology*, can be grouped together, as they are used with similar purpose. Firstly, they indicate the worldwide impact of the fields, and worldwide concern and engagement with them. This worldwide impact is written about with varying positive and negative associations- evoking both utopian and dystopian imagery.

In terms of positive associations, many of the contributions of *international* are due to articles which discuss the International Centre for Genetic Engineering and Biotechnology, a centre which focuses on use of Genetic Engineering to benefit developing countries. This shows progressive benefits of genetic engineering being emphasised and discussed. In addition, the influence of iGEM in increasing discussions about synthetic biology is clear when analysing *international* as a collocate of *synthetic biology*- most of the articles which feature this word relate to explaining iGEM. This is helping to promote synthetic biology as a global community, and increasing the coverage of synthetic biology in the media- a positive influence in terms of *international* being a collocate of *synthetic biology*.

While there are positive associations of *international*, the KWIC lines also indicate some associations which are more negative and critical of the fields. Many of these relate to regulation in genetic engineering and synthetic biology. Questions of whether an international ‘body’ or ‘framework’ is necessary for controlling genetic engineering and synthetic biology are frequently asked in the articles which feature in the corpus. There is obvious public concern over whether there should be internationally equated regulation, with use of synthetic biology controlled in the same way across the world, to prevent discrepancies of use and transfer across borders. Negative influence is also clear when analysing *global* as a collocate. In one news story, published by the Financial Times, synthetic

biology is talked about in the context of it creating a 'global pandemic': '[r]eflect upon the potential for a global pandemic created by synthetic biology [...] what are the international means through which governments will negotiate their responses to such an event?' (Stavridis 2014). Using this world transfers existing associations that readers have with a word- in this case negative ones, as a pandemic is by definition the worldwide occurrence of a dangerous disease- onto a new field they do not possess as much knowledge about. The impression created is that synthetic biology is a disease that is going to spread with dangerous consequences if left unchecked.

Although hyperbole is commonly made use of in the media, both to entertain and as a rhetorical device, it is salient in the discussion of genetic engineering and synthetic biology. The fact *extreme* and *nuclear* (as in nuclear bomb) feature as collocates of genetic engineering exemplifies this.


One article from *The Atlantic* uses a quote from Friends of the Earth, who describe synthetic biology as 'an "extreme form" of genetic engineering' (Garthwaite 2014). In this context, *extreme* carries negative semantic prosody. Using the NOW corpus to search for the collocates of *extreme* verifies this. A selection of the top collocates of *extreme* are *poverty*, *violence*, *pain* and *drought*- all inherently negative words. As *extreme* is frequently used to describe negative things, it can be used in a context of negative association, which is the effect created when it is used to describe genetic engineering and synthetic biology.

An article entitled *The Ethical Use of Big Data* talks about technologies and the dual-use debate. Genetic engineering is used along with nuclear energy as an example for the reader. '[m]ost breakthrough technologies have dual uses. Think of atomic energy and the nuclear bomb or genetic engineering and biological weapons' (Barabási 2013). The writer goes on to say '[t]he model we scientists follow is simple: We need to be transparent about the potential use and misuse of our trade'. Promoting transparency to address the dual-use debate is progressive, and further emphasises the need to be transparent in language use when addressing the public.

With both genetic engineering and synthetic biology, the collocates show that a lot of media discussion centres on them as 'new' disciplines. However, there is a difference in how this is done for each of the terms. While a word like *new*-the clearest and most simple definer of a new field- features as a collocate for both *genetic engineering* and *synthetic biology*, there are additional words which feature for one or the other. These words are also used to describe the fields as new, but do so in different ways, hence revealing different attitudes towards them. For example, genetic engineering is discussed as a new discipline by means of contrast. The words *conventional* and *traditional* are used to do this: genetic engineering is portrayed as an alternative to different methods that are considered more







traditional. For example, genetic engineering is frequently juxtaposed in opposition to conventional breeding techniques: an article published on *Crosscut* states '[u]nlike conventional breeding, genetic engineering invites crossbreeding of unrelated species' (ODonnel 2013). In a different approach, when writing about synthetic biology, a more forward thinking, progressive impression is created. The feature of *emerging*, *evolutionary*, and *cutting-edge* as collocates of *synthetic biology* illustrate this. These words suggest exciting advancement is in progress (top collocates of *cutting-edge* include *technology*, *innovation* and *solutions*). Additionally, many of the hits for *new* as a collocate of *synthetic biology* are from the use of the phrase 'brave new world'. One article from *The Independent Online* states '[t]he brave new world of synthetic biology has taken a major step forward as the UK opened its first DNA factory - manufacturing genetic components that will be used to tackle everything from global warming to vaccines' (Bawden 2016). 'Brave new world' is a common idiom, making unfamiliarity with the field of synthetic biology more relatable, and adding an element of mystery and apprehension.

The use of language in discussion of genetic engineering and synthetic biology is often characteristically tentative. The word *possible* collocates with *genetic engineering*, and the word *potential* collocates with *synthetic biology*. These words both function as hedges, lessening the impact of the claims made in the articles and indicating uncertainty about the effects of the fields and the work emerging from them.

Analysing the collocates of synthetic biology and genetic engineering reveals that both positive and negative attitudes are held towards these fields. This is information which can be used; the attitudes should not be diminished, but it should be acknowledged why they may be held, and how the positive can be accentuated over the negative. Considering semantic prosody of the words used to describe the fields can help with this. The collocates also teach that for work in synthetic biology to have a solid impact, it is important not to add to the already uncertain public reception that is held by using hedges when communicating the work.

### 3.2. GMO

When starting to research the attitudes towards genetically modified organisms, both *genetically modified organism* and *GMO* were used as search terms. *GMO* has been focused on because it has more widespread use- it is a commonly known and used acronym. This means there were more collocates of *GMO* with higher frequencies, which are therefore more significant.





Ranking of Collocate	Collocate
1	Modified
2	Safe
3	Free
4	Other
5	New
6	Organic
7	Mandatory
8	Scientific
9	Public
10	Genetically-Modified
11	Human
12	Environmental
13	Growing
14	Agricultural
15	Global
16	Commercial
17	Genetic
18	Good
19	Dangerous
20	European

Figure 5: Top 20 adjectives which collocate with *GMO*, in NOW corpus (Davies 2013).


### 3.2.1 Analysis of *GMO* Collocates

Figure 5 shows the top twenty adjective collocates of *GMO*. Again, by using the KWIC lines, further conclusions can be drawn about attitudes towards GMOs from the language used to discuss them.

*Genetically-Modified* and *modified* occur as collocates because of the frequency of explaining what *GMO* stands for. In some cases, this is used as more of a rhetorical device than an educational one, to make the phrase and the concept stand out and be noticed, by repetition: '[t]he issue was genetically modified organisms, or *GMOs* as they're often known in the food industry' (McAuliff 2014).

The majority discussion about *GMOs* is in the context of *GMO* foods. While this is understandable, as food production and consumption is a global issue that directly affects everybody, it indicates that people may not be aware of the full extent of what *GMOs* are, and what they can be used for. Here, the science community is, understandably, concerned with reporting the issues that are currently affecting the largest number of people, but a large proportion of the coverage *GMOs* get is due to the controversy associated with their use in foods.

There are collocates which, when examined further, evidence the atmosphere of distrust and controversy, and voicing of safety concerns that surround *GMOs*: *free*, *safe*, *new*, *mandatory*, *public*, *human*, and *environmental* are key examples. *Mandatory*, *public*, *human* and *environmental* all collocate because of articles which talk about public health concern relating to *GMO* foods. There are



many debates which are discussed, such as whether mandatory labelling of GMO foods should be enforced, and what the real effects on public health are.

A lot of the language used in discussion of GMOs is conditional: for example, *whether* often occurs next to the collocates of *GMO*- ‘whether GMOs are safe’. This indicates the uncertainty and debate that surrounds their use.

### 3.3. *Biosensor*

*Biosensor* has been chosen as a final search term because of the significance for the Newcastle iGEM project. Hence, by searching in for this term in corpora, information can be gathered that can be used to aid the communication of our project, on a more specific level.

Firstly, it must be noted that all of the collocates of *biosensor* have relatively low frequency when compared to the frequencies of the collocates of *genetic engineering*, *synthetic biology*, and *GMO*. This is due to the lower frequency of occurrence of *biosensor*: this is a less widely discussed topic.

#### 3.3.1 Analysis of *Biosensor* collocates

Figure 6 displays the top 20 adjectives which collocate with *biosensor*.

Ranking of Collocate	Collocate
1	New
2	Wearable
3	Sensitive
4	Optical
5	Global
6	Small
7	Electrochemical
8	Portable
9	Rapid
10	Early
11	Other
12	Simple
13	Conventional
14	Analytical
15	Accurate
16	Non-invasive
17	Photonic
18	Proprietary
19	Single-use
20	Tiny

Figure 6: Top 20 adjectives which collocate with *Biosensor*, in NOW corpus (Davies 2013).

*New* is the word which most frequently collocates with *biosensor*. Many of the articles which discuss biosensors do so in the light of creating a new technology, with biosensors providing a new method of detecting and measuring different things. Using *new* creates excitement about the noun the adjective is modifying, *biosensor*, encouraging readers by showing it is innovative technology they will want to find out about. Along with excitement, newness also comes with apprehension: new things are, by definition, not established practises. In order to build trust in the Newcastle iGEM project, newness is an element which must be treated positively. The development platform is a new way of creating biosensors, but is informed by adapting existing principles and practises to improve the efficiency of a process.

Many of the collocates of *biosensor* illustrate how discussion around biosensors uses key features which describe and define them. Examples of these collocates are *sensitive*, *small*, *portable*, *rapid*, *simple*, *analytical*, *accurate*, and *tiny*. These adjectives all portray positive, convenient attributes of biosensors, using language which is simple to understand to describe the technology. When writing for an audience that may not know what a biosensor is or does, this type of language makes it clear. On the other hand, there are also collocates which show that jargon- more technical and field specific language- is also used when discussing biosensors. For example, *optical*, *electrochemical*, *photonic*, and *proprietary* indicate articles that use scientific discourse, and can assume the knowledge of their reader to reach the understanding of these terms. This is where audience awareness is particularly important when communicating science: the knowledge level of the audience you are writing for will dictate the detail used in using and defining these terms. For example, an article published by *Digital Journal* in 2016 analysing the biosensor market uses the terms *electrochemical* and *optical* in their breakdown of the type of technology the biosensor utilises. This is report targeted at those working in the biosensor industry.


When promoting and explaining Sensynova, the Newcastle University 2017 iGEM project, by making use of a knowledge of how biosensors are currently discussed, it can be ensured that our dialogue fits in with the current dialogue, and makes use of key defining features.

#### 4. Conclusion

By using corpus linguistics to gain an awareness of how genetic engineering, synthetic biology, GMOs and biosensors are discussed in popular media, two major goals have been accomplished. Firstly, an understanding of how much information is currently discussed in the media, and the attitudes which are portrayed towards these technologies, has been gained. Secondly, an insight into language use when communicating these areas of science has been gained. Overall, this is information which we aim to







use to produce a set of advisory guidelines for communicating science. Furthermore, the insight gained will be used to inform additional public engagement activities during the iGEM project, including talks with stakeholders.

A summary of the main points to carry forward that have been identified, and plans to implement them, is as follows:


- 1) In comparison to genetic engineering, synthetic biology has very low coverage in the media. In order to increase awareness of synthetic biology, additional platforms could be used to reach the public audience. By writing a blog about education and public engagement activities throughout our iGEM project, we aim to engage a wide audience with synthetic biology.
- 2) Key terms used in the discussion surrounding synthetic biology have both positive and negative associations, and portray both positive and negative attitudes towards the field. It is important to acknowledge this when discussing our project, and make use of the knowledge of the effect existing semantic prosody of adjectives can have if they are used in descriptions.
- 3) Regulation of synthetic biology is one of the main concerns that features in the discussion of synthetic biology. We are ensuring we consider regulation more, and assess how our project fits within current regulation.
- 4) An understanding of linguistic devices used in discussion of these topics can help increase the clarity and success of scientific communication. Examples identified in this research include the use of hedges and the use of conditional modals. These can relate to uncertainty in language use. When these effects are created by those promoting a technology, distrust in its potential users can be fuelled, so we should minimise this in communicating the project.
- 5) Transparency and audience awareness (for example, will your audience understand jargon) is often key in gaining trust and engagement with a project. Fostering transparency in language use can aid transparent communication, so many of the points in our guidelines for communication of science in iGEM will focus on this.

Finally, this is work which can be built upon in iGEM in years to come. It is clear from the research that iGEM has already made a big impact in terms of increasing the discussion around synthetic biology. As the corpora used in this research, NOW and COCA, continue to be updated, and discussion around these topics continues, the searches can be repeated. It would then be possible compare search results as time progresses to view how attitudes and coverage of the topics advance, and think further about what can be done to improve communication and engagement.



## List of References

- Ahmadian, Moussa, Hooshang Yazdani and Ali Darabi. 2011. 'Assessing English Learners' Knowledge of Semantic Prosody through a Corpus-Driven Design of Semantic Prosody Test'. *English Language Teaching* 4(4): 288-298.
- Barabási, A. 2013. 'The Ethical Use of Big Data'. Last Accessed 18<sup>th</sup> August 2017, from: <http://www.politico.com/story/2013/09/scientists-must-spearhead-ethical-use-of-big-data-097578>
- Bawden, Tom. 2016. 'DNA factory unveiled at UK university'. Last accessed 18<sup>th</sup> August 2017, from: [https://www.iol.co.za/news/dna-factory-unveiled-at-uk-university\\_2006383](https://www.iol.co.za/news/dna-factory-unveiled-at-uk-university_2006383)
- Davies, Mark. (2008-) *The Corpus of Contemporary American English (COCA): 520 million words, 1990-present*. Available online at <https://corpus.byu.edu/coca/>.
- Davies, Mark. (2013) *Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day*. Available online at <https://corpus.byu.edu/now/>.
- Digital Journal. 2016. 'Biosensors Market: Global Industry Analysis Report, Share, Size, Growth, Price Trends and Forecast, 2024: Fractovia.org'. Last Accessed 22<sup>nd</sup> August 2017, from: <http://www.digitaljournal.com/pr/3149850>
- Evison, Jane. 2010. 'What are the basics of analysing a corpus?' in Anne O'Keeffe and Michael McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*, London: Routledge. 122-135.
- Flowerdew, John. 2013. *Discourse in English Language Education*. New York: Routledge.
- Garthwaite, Josie. 2014. 'Beyond GMOs: The Rise of Synthetic Biology'. Last Accessed 20<sup>th</sup> August 2017, from: <https://www.theatlantic.com/technology/archive/2014/09/beyondgmoss-the-rise-of-synthetic-biology/380770/>
- Louw, B. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair*. Amsterdam: John Benjamins. pp. 157-176.
- McAuliff, Michael. 2014. 'Americans Are Too Stupid For GMO Labeling, Congressional Panel Says'. Last Accessed 19<sup>th</sup> August 2017, from: [http://www.huffingtonpost.com/2014/07/10/gmolabels\\_congress\\_n\\_5576255.html](http://www.huffingtonpost.com/2014/07/10/gmolabels_congress_n_5576255.html)
- ODonnel. 2013. 'The ABCs of GMOs'. Last Accessed 20<sup>th</sup> August 2017, from: <http://crosscut.com/2013/10/primer-on-gmos-i522-odonnell/>



Stavridis, J. 2014. 'The dawning of the age of biology'. Last accessed 20<sup>th</sup> August 2017, from:  
<https://www.ft.com/content/36218738-6355-11e3-a87d-00144feabdc0>

Wood, Linda, and Rolf Kroger. 2000. *Doing Discourse Analysis: Methods for Studying Action in Talk and Text*. London: Sage Publications.

### Acknowledgments



We are grateful that this project was supported by NUHRI's (Newcastle University Humanities Research Institute) Challenge Labs Scheme [Summer17].