LINGUISTIC WEB SCIENCE:

LOOKING FOR MEANING ON THE WORLD-WIDE-WEB

I

Language allows us to give an identifiable form and substance to our most creative impulses. As often happens when talking of such things, Shakespeare got there first and said it best, this time in *A Midsummer Night's Dream*: "*And as imagination bodies forth the forms of things unknown the poet's pen turns them to shapes, and gives to airy nothing a local habitation and a name*". Our focus in this course is linguistic creativity, not quite the kind we associate with Shakespeare, but the more day-to-day variety we encounter in casual conversations, jokes, cartoons, comics, newspaper columns, movies and web postings. This course should be viewed neither as a guide book nor a history book, but as a highly-opinionated (and, hopefully, mildly diverting) tour of how words and phrases are used to achieve novelty, variety and concision in everyday language, and especially in the language of the World Wide Web.

Why this special emphasis on the web? Computers have influenced our relationship with language in many ways, some remarkable and some mundane, but few developments have been as significant as Tim Berners-Lee's invention of the World Wide Web. As recently as the early 1990s, if you wanted to dig up a definition, seek out a synonym or query a quotation, your first port of call would have been the nearest bookshelf, sagging under the weight of hefty tomes like the *Oxford English Dictionary* and *Bartlett's Familiar Quotations*. Failing that, you might actually have paid a visit to the reference section of a real library. The web has changed all of that, and the door-to-door encyclopaedia salesman has gone the way of the muffin man, the chimneysweep and the rag-and-bone cart. It's not just that the web now provides all of these reference works online and mostly free of charge, with easily queried interfaces and rapid response times. It's much more significant than that, for the web has become the ultimate linguistic resource in its own right, big as life and just as surprising. As web technologies become increasingly pervasive, and the web itself becomes the medium of choice for the

expression of popular culture, the changes wrought on language have accelerated in pace. For good or ill, the web has become an incubator for rapid innovation in language.

Let's suppose you want to know the official state nickname of New Jersey. Each American state has its own definitive tagline, from *The Sunshine State* (Florida) to *The Lone Star State* (Texas), and many an authoritative resource, whether in print or on the web (Wikipedia, for example), can tell you that New Jersey is officially known as *The Garden State*. Unchanging factoids such as these are perfect for reference books, of the kind handed down in families from one generation to the next. But suppose you instead want to find an *un*official nickname for New Jersey, a more topical tagline that represents how speakers actually feel about the state right now. Then those books on the shelf are already out of date, and even dynamic sites like Wikipedia are unlikely to contain anything as unconventional and lacking in authoritativeness as a new up-and-coming linguistic expression. But you can also turn to the web as a whole, and use a search engine like Google or Bing to unearth recent coinages that are still young and fresh and still largely outside the mainstream.

For instance, Google allows you to formulate queries like "*The * State: New Jersey*" where the * is a wildcard that will match any term in that position. One of the matches we find on the first page of results is for a book called *The Soprano State: New Jersey's Culture Of Corruption* by Bob Ingle and Sandy McClure. The book borrows its name from the TV series *The Sopranos*, a drama that centres on a fictional New Jersey mob boss, Tony Soprano, and his crew of tragicomic villains. Using the web to further determine the extent to which this unofficial moniker has been adopted by writers and pundits, we find that even *The Economist* newspaper now playfully refers to New Jersey as *The Soprano State*. While the name was coined *off*-web, in the title of a book with real pages and hard covers, it gained currency as a creative expression *on* the web. Of course, *The Garden State* is still the undefeated champ when in comes to describing New Jersey, with over four million references on the web, but *The Soprano State* is the fresh new face of the Zeitgeist, with a substantial body of 300,000 web hits and counting.

Every linguistic innovation has to originate somewhere, yet the speed at which new coinages now become part of the measurable currency of the web is startling. When

formed as part of original web content, these new expressions can yield search results on the very same day they are coined. As we get more and more of our news and opinion from the web, we are increasingly likely to be present at the birth of a resonant new coinage. Whenever a new phrase does catch our eye, we can immediately check its provenance with a quick web search. Here's just one example. Responding in April 2010 to the second televised leaders' debate in the British general election, Marina Hyde's article for *The Guardian* newspaper focused on the media's coverage of the debate, and on the spin placed on the event by various "*spindroids*" – Hyde's novel term for a politico capable only of pre-programmed, robotic responses (compare this with the more conventional *spin doctor*, and Dan Quayle's immortal *Dr. Spin*). Though writing in haste on the night of the debate for the next morning's edition, Hyde also floated three other memorable coinages in her article: the lexical invention "*arsesoisie*" (presumably a blend of "arse" and "bourgeoisie") and the eye-catching phrases "*weapons-grade wisdom*" and "*the live abortion of democracy*". The latter was quickly echoed across the web, and within a fortnight had appeared in over 5,000 web documents that correctly attributed the phrase to either Hyde or *The Guardian*. Whether used literally or metaphorically, "abortion" is a hot-button topic on both sides of the political divide. However, the coinage "*weapons-grade wisdom*" proved to be far less popular, and in the following weeks was picked up just 3 times on the web. Some fledgling phrases never take flight.

Curiously, "*weapons-grade wisdom*" had first been coined in a little-visited blog a year earlier, but that innovation also failed to set the blogosphere alight. The phrase would not be used again (on the web, at least) until Hyde's re-invention in 2010, which was most likely a case of independent discovery. Nonetheless, in the words of creativity theorist Margaret Boden, Hyde's invention was merely psychologically creative, or *P-Creative*, while the original 2009 coinage was also historically creative, or *H-Creative*. In language we often strive for *H-Creativity*, but unless you are a professional wordsmith, our communicative purposes are just as well served by good *P-Creative* craftsmanship. Though we can be disciplined in our approach, linguistic creativity is still more an art than a science, so who can say why "*weapons-grade wisdom*" never took off? The phrase is clearly a creative riff on the more familiar "*weapons-grade uranium*" (which yields over four million web hits on Google), and the "om" of "wisdom" even rhymes with the

"um" of the "uranium" it replaces. But some variations grab the imagination more than others. The phrase "*weapons grade bolognium*" – which has much the same negative meaning as Hyde's ironically radioactive "*wisdom*" – has made its way into thousands of web articles since its first use on the animated series *Futurama* in 2000. Perhaps "wisdom" just doesn't sound enough like a chemical element, unlike the cod-science "*bolognium*", "*unobtainium*" (from the films *Avatar* and *The Core*), "*adamantium*" (from the *X-Men* movies and comics) and "*bolonium*" (once used in *The Simpsons* TV show)?

II

Hyde's "*weapons grade wisdom*" may prove more popular third-time out, for the factors that determine whether a new coinage will be widely adopted have as much to do with culture as language. Our primary goal in this course is an exploration of how language supports creativity, which will lead us to examine the linguistic factors that shape the development of new words like *bolognium* and *spindroid*, as well as creative riffs on familiar phrases like "*weapons grade bolognium*" and "*weapons grade wisdom*". Where possible, we'll consider the cultural backdrop to these innovations as well, but one should never confuse a just-so story about popular culture with a real explanation for anything. So the philosophy that suffuses this course is essentially a computational one: our discussion of processes and representations follows from the belief that the best way to understand creativity in humans is to speculate on what would be needed to model the same kind of productive behaviour on a computer. The computational perspective is a revealing one whether or not you share the belief that computers can be intelligent *and* creative. Cold-blooded mechanisms that they are, computers have no self-image to preserve and no time at all for the cultural pieties that surround human creativity.

Despite advances in AI, computers are still making very modest progress in the realms of creative endeavour. Computers can find new ways of solving problems in mathematics, and are capable of generating rather good musical variations (in both the jazz and classical traditions), and acceptable but so-so art. In fact, researchers have developed algorithmic systems for generating all of the following: mediocre (but steadily improving) representational art scenes; unspirited but mathematically interesting abstract

art; automated stories that show an appealing naïveté whenever they wonder outside the strict confines of their knowledge bases; humorously original, if not hugely sophisticated, puns and riddles; clever acronyms with a twist of sarcasm; weak jokes of a particularly formulaic bent; and automated poetry that would make a Vogon blush with embarrassment. Each development is promising in its own way, for this kind of work has to start somewhere, and each new baby-step takes researchers incrementally closer to banishing the hoary old myths that surround creativity. Indeed, we may yet broaden our conception of creativity to the point that it is no longer seen as a uniquely human quality.

Ultimately, there is no place to hide with the computational approach, and its incisiveness as an analytical tool really does cut both ways. So it will become abundantly clear that the various processes and representations described in this course fall short of a complete picture of any aspect of creativity. Yet the beauty of the computational approach is that even an incomplete picture has practical benefits that researchers can continue to build upon. Computers are excellent tools for automated brainstorming on an almost industrial scale, and as we'll see, they can be used to road-test different ideas with large-scale simulations, both to compare the productivity of different creative strategies, and to see the fullest range of outputs that a given strategy can deliver. A happy consequence, then, of a computational perspective is that we often end up with useful, if limited, tools that others can use. A variety of tools for supporting linguistic creativity have been constructed on the basis of the simulations discussed in this course. Readers are encouraged to visit *Afflatus.UCD.ie* to play with these tools for themselves.

III

In the course of our exploration we'll discuss everything from clichés and stereotypes to jokes, similes, metaphors and analogies. We'll introduce and explain linguistic terms like exploitations, optimal innovations, snowclones, scripts, blends and XYZ constructs. We'll make frequent references throughout to the language of the web, and at these points readers are strongly encouraged to open up a web browser, fire up a favorite search engine, and start exploring for themselves. We'll occasionally go to town in our analysis of language on the web, and use computer programs that are faster and less easily bored

to automatically trawl large tracts of the web for us. Because the web can be seen as a vast corpus of contemporary language use, it presents a wonderful opportunity to employ some heavy-duty crunching of words and numbers. This will allow us estimate the creative range of different kinds of linguistic device, from similes (ironic and otherwise) to blended word forms and XYZ metaphors.

In fact, our analyses will do more than provide evidence for the seething diversity of creative language on the web. By casting our automated net far and wide over the texts of the web, we'll show that a relatively small set of commonplace but very useful linguistic constructions is responsible for a staggering variety of creative expressions. These forms are, in effect, support structures for linguistic creativity. Such support structures are used by creative professionals and amateurs alike, from polished comedians and accomplished literary stylists to ranting bloggers and thumbsore twitterers. Ironically, it seems, convention-defying creativity often follows a pattern, and it pays to know how and when to exploit the most common strategies and structures for crafting a meaning-rich creative expression. But we'll encounter the *hows* and *whys* of linguistic creativity soon enough. Before we get into full swing, let's take a moment to consider the topics that lie ahead.

IV

Concerns for the linguistic coherence of the web and the intelligence of its users are not entirely without foundation. Many times it can seem that the web is not so much a bubbling cauldron of many voices and even more creative expressions, but an echo chamber in which the silliest, the most inane and the most extreme views reverberate the loudest. In fact, if one wanders off the main thoroughfares of the web, with their trusted shop fronts and authoritative information points, one is likely to encounter a great deal of nonsense, both conceptual and linguistic, in its back alleys and basements. However, we should remember that this undesirable content is all too easy to find precisely because the web makes *everything* easy to find, good or bad. To paraphrase the sci-fi writer Theodore Sturgeon, whose defence of the quality/crud ratio in science fiction writing is often called *Sturgeon's Law*, 90% of the web's content is crud "because 90% of *everything* is crud".

From the perspective of creative language, there is much to love about the web. One simply needs to exercise care about where one roams and what one is willing to believe. If used thoughtfully, in ways we shall discuss in this course, the web in its entirety can be viewed as a vast and continuously growing creative thesaurus, not like the reference books we find on our shelves, and not like the specific reference sites we find as distinct *parts* of the web, but an almost unbounded representation of what is possible in language. To be creative with language, one must explore the regions that lie between what is conventional and what is possible. The web enables this kind of linguistic exploration on the grandest scale imaginable. Who can say what Shakespeare might have thought of the web? But if ever a fairy toy allowed us to distil linguistic magic from the seething brains of lovers and madmen, the web is it.

T. Veale,

Dublin, August 2011