Quaderns de Filologia. Estudis lingüístics. Vol. XVIII (2013) 97-113

WORD-FORMATION AND WORD-CREATION: A DATA-DRIVEN EXPLORATION OF INVENTIVENESS IN NEOLOGISMS

Pierre J. L. Arnaud
CRTT, Université Lumière, Université de Lyon

1. INTRODUCTION

When we come upon occurrences of recently coined lexemes for the first time, some strike us by their inventiveness while others may well pass unnoticed[1]. When we say that a word is inventive, however, this is obviously a case of process-for-result metonymy: what is inventive is the forging of that lexeme, not the lexeme *per se*, and this should be kept in mind in what follows. This feeling of inventiveness is an intriguing aspect of lexicogenesis and is worth exploring.

The lexemes *inventive* and *inventiveness* are not present as entries in handbooks of morphology and lexicology, evidence that they do not constitute terms of those domains, and, to the best of my knowledge, the fact that some words appear as inventive has not been explored. The semantically close lexeme *creativity*[2], however, is clearly a morphological term. It frequently occurs in contexts where it is opposed to productivity, and this opposition is more or less identical with that between word-creation and word-formation, non-rule-governed and rule-governed lexicogenesis, and also extra-grammatical and grammatical morphology (see for instance Marchand, 1969: 2-3; Bauer, 1983: 63; 2001: 64; Ronneberger-Sibold, 2010; Booij, 2012: 20-22; Mattiello, 2013). *Creativity* therefore denotes the production of new words without recourse to grammatical rules, whereas *inventiveness* is suggested in this article as the name of a subjective property of words.

Since *inventiveness* is not currently a morphological or lexicological term with a clear definition, and keeping in mind that its terminological denotation should not overlap that of the already existing *creativity*, we can start from

[1] I am indebted to Dr. Laurent Arnaud for statistical help.

[2] *Creativity* figures in the definition of *inventive* in the *Concise Oxford English Dictionary* (2011).

its general meaning. If we focus on word-formation, possible components of inventiveness are:

(1) unexpectedness, i.e. the form-meaning relationship is indirect;
astuteness, i.e. a complex concept is named aptly;
compactness (terseness, economy of means);
playfulness, i.e. the users detect humour in the formation;
live metaphor and metonymy (esp. if far-fetched or colourful);
need for an interpretive effort.

Some of these components are not entirely independent: for instance, an indirect form-meaning relation will require more interpretive effort than a direct one. A serious problem is that most of these qualities are subjective and hardly operationalizable experimentally. This is because inventiveness is a construct, i.e. a psychological variable about whose existence there can be commonsense agreement, but which cannot be defined precisely and measured. Given this situation, one exploratory solution to investigating this construct consists in using the subjectivity of a group of informants exposed to a set of neologisms.

2. FIRST INVESTIGATION

2.1. *Method*

It was decided that informants' subjectivity could be most practically tapped by means of a questionnaire asking them to grade neologisms with reference to their inventiveness. The first stage consisted in gathering the stimulus units. Web sites of the *Word of the Year* kind could not be used because they contain neologisms that are topical and tend to be on the playful or sarcastic sides, and a broader spectrum was needed for the investigation. Dictionaries of neologisms, in particular Algeo (1993) and Barnhart *et alii* (1990), were therefore preferred. The units were drawn at random (the 5th word of every 5th page, etc.) until 100 units were collected. Using a table of random numbers, the list was then reduced to 40 units, a number which seemed acceptable for inclusion in the questionnaire if definitions were added. As the items result from a random draw, there was no control for word class or lexicogenetic type. The units were searched in the *Oxford English Dictionary* on-line for dates of attestation: 6 were absent from the *OED*, 2 appeared in the late-19th century, 6 between 1900 and 1950, and 26 are dated post-1950. To select the relevant senses in cases of polysemy, and also in order to reduce the risk that

some recent neologisms might be unknown to the informants, dictionary-style definitions like the following example were written:

(2) cryonics = the deep-cold conservation of human bodies in the hope that they can be resuscitated in the future.

The following instructions were placed at the top of the page:

(3) For each item, please tick a box corresponding to the degree of inventiveness you think played a role in the formation of the word. Use your spontaneous idea of inventiveness.
0 = not inventive at all
1 = not much inventiveness
2 = rather inventive
3 = highly inventive

The four possible grades were used in order to preclude a possible bias towards a central answer. Item scores in the 0 to 3 range therefore resulted for each neologism in the test. In order to counterbalance questionnaire fatigue, two versions were printed with the items in different random orders (one of the versions is reproduced as Appendix 1). The questionnaires were handed out by intermediaries to native speakers, some of whom were British teachers of English to non-natives, others American students present in Lyon. The informants were told to fill the questionnaires at their own convenience. Twelve copies were returned.

Initial examination showed that some informants had been more generous or more severe in their marking than the others, so item scores were standardized. Inter-rated agreement was determined using Kendall's W. Its low value, $W = .309$ ($n = 12$), shows that judgments were far from unanimous. Obviously, individuals react differently to such verbal stimuli.

The average of the z-scores attributed to each unit was then calculated, and the units were rank-ordered accordingly. Using the ten most and the ten least "inventive" units (presented in Table 1) allows us to maximize the contrast. In order to compare the two sets of ten items in a reliable fashion, the significance of the difference between them was established, and for this purpose the total numbers of "0", "1", "2" and "3" scores were calculated for the two sets of items. The corresponding χ^2 value of 91.7 is significant (3 d.f., $p < 0.01$), so, in spite of the low degree of inter-rater agreement, the comparison can be carried out on firm grounds.

TABLE 1: Units with the ten highest and lowest inventiveness scores (1st investigation)

	unit	z
1	hit man	1.02
2	smoothie	0.92
3	moonlighter	0.86
4	quark	0.76
5	bullet train	0.73
6	to amp up	0.71
7	baby boom	0.69
8	egghead	0.68
9	hang-up	0.67
10	ballute	0.54
…	…	
31	glide bomb	-0.43
32	multihull	-0.47
33	ozone hole	-0.51
34	extravehicular	-0.58
35	systems analyst	-0.60
36	condo	-0.61
37	profit centre	-0.67
38	to pressure	-0.75
39	pollee	-0.77
40	inflatable	-1.15

2.2. *Discussion*

In the following discussion, the term *transparent* will be used for units whose form leads effortlessly to their meaning. For instance, knowing the base of a derivative, the addition of a transparent suffix leads to a predictable meaning, as in *sing* > *singer* ("one who sings"). The word judged to be the least inventive, *inflatable*, is a perfectly transparent derivative. This is also the case for the second-lowest, *pollee*. We find two more derivatives, *extravehicular* and *multihull*, among the ten low-inventiveness words. The complexity of

extravehicular, which is formed around a bound root, or the metonymy present in *multihull* (a characteristic for the entity) did not make them appear more inventive.

Stricto sensu, $[NN]_N$ compounds are never transparent, since the modification relation is never explicit. However, some relations are statistically predictable (see for instance Gagné & Spalding, 2009), like the contents-container relation in *wine bottle* and in other names of containers with *box*, *closet*, etc. as heads. In the case of compounds, the term *transparent* will refer to the predictability of the N1 to N2 relation. The four non-inventive compounds, *profit centre*, *systems analyst*, *ozone hole* and *glide bomb*, are relatively transparent (for instance, a system analyst analyzes systems), with the possible exception of *ozone hole*, as compounds with N2 *hole* include several possible relations and metaphors (cf. *rabbit hole*, *bullet hole*, *pigeon hole*, *memory hole*, *asshole*). However, since the meaning was given in the accompanying definition, the relation certainly appeared as uncomplicated. The other two non-inventive units are a clipping, *condo*, and a N > V conversion, *to pressure*.

Two derivatives, *moonlighter* and *smoothie*, figure among the ten most inventive units. The inventiveness score of *moonlighter* might appear as problematic at first sight, as it is a semantically predictable agent name of *moonlight*; however, *moonlight* is a verb resulting from conversion and including a metonymy (the name of an entity present during an activity for the activity); the impression of inventiveness therefore probably comes more from the base than the derivative itself. *Smoothie* has the *-ie* diminutive suffix as in *sweetie*, *dearie*, used to mark familiarity. Denoting a drink, it includes a "name of a characteristic for the entity" metonymy.

The inventive compounds are *hit man, bullet train, baby boom*, and *egghead*. *Hit man*, the unit with the highest average score, seems relatively transparent (agentive, "a man who hits", as in *workman, con man*, etc.), but it also compacts encyclopedic knowledge about crime: *hit man* means much more than "a man who hits", and it is a highly terse unit. In *bullet train* we have a live *in praesentia* metaphor based on shape and speed, and in *egghead* one based on a purported link between being an intellectual and skull shape and fullness, so a rather complex figurative meaning. *Baby boom* includes a metaphtonymy on its N2 (the noise for an explosion; an explosion for a rise), like *oil boom*, *ethanol boom*, *mining boom*, *property boom*, but also, and in contrast to these examples, its N1 denotes human beings, not commodities or activities, so the pattern analogy is only partial.

Ballute is a blend, whose source words *balloon* and *parachute* can be recognized with some effort. *To amp up* is a clipping leaving one syllable out of three, presumably with a conversion from the noun *amplifier*, otherwise the

particle would be redundant. *Hang up*, an informal V+particle noun dated 1959, apparently from the notion of being suspended, is therefore metaphorical, but the metaphor is not easily accessible. *Quark* results from an arbitrary naming act, so from word manufacture (Bauer, 1983: 239), and as such is opaque.

2.3. *Conclusion*

First of all, as noted above, in spite of a low degree of inter-rater agreement, a significant difference between the first and the last ten words in the sample appeared when they were rank-ordered. What we see in this rank-ordering of the units in terms of perceived inventiveness is that although derivatives and compounds are found both in the high and low ranking groups, there is overall a difference in their apparent transparency, the "inventive" units requiring more interpretive effort. Tropes are more characteristic of the high inventiveness group and, although isolated examples invite caution, one finds a blend and a manufactured word in the high inventiveness group and none in the low inventiveness group.

3. SECOND INVESTIGATION

3.1. *Method*

In view of the results of the first investigation, it was decided to use another questionnaire based on the same principles, but with an organised list of neologies instead of a random one in order to allow differences to appear more clearly. The units were gathered from the same dictionaries as in the first investigation, with supplements from the *Wordspy* website[3]. The following categories were targeted (examples for the categories can be found in Appendix 2):

(4) simplex words (literal)
simplex words (with tropes)
derivatives (literal)
derivatives (with tropes)
compounds (literal)
compounds (with tropes)
conversions
back-formations
clippings

[3] <http://www.wordspy.com/>; last consulted on 6 January 2013.

blends
units formed with splinters
various

The number of items was raised to 48, i.e. 4 in each of the 12 categories. The items with their definitions were included in the questionnaires in two different random orders, and the same instructions and grading system were retained (one version of the questionnaire is reproduced as Appendix 3). Some of the informants had already taken part in the first investigation, but more students returned forms and the number of questionnaires returned was 17. Data treatment was similar to that of the first investigation. Inter-rater agreement, as measured by Kendall's W, was .317 (n = 17), a low figure as earlier. Table 2 includes the first and last ten items rank-ordered for z-scores averaged across informants. The significance of the difference in the scores attributed to the two groups taken together was tested in the same way as for the first investigation, and the difference was here again significant ($\chi^2 = 164.246$; 3 d.f.; $p < .001$).

TABLE 2: Units with the ten highest and lowest inventiveness scores (2nd investigation)

	unit	z
1	glitterati	1.09
2	hacktivist	1.05
3	snowmaggedon	0.98
4	commuterdom	0.96
5	poppers	0.89
6	newbie	0.46
7	nerd	0.42
8	raunch	0.40
9	prequel	0.38
10	klick	0.37
…	…	
39	reader	-0.45
40	celeb	-0.65
41	mountain bike	-0.72
42	to bartend	-0.79

	unit	z
43	to curate	-0.83
44	environmentalist	-0.84
45	to waitress	-0.86
46	bird flu	-1.07
47	diet pill	-1.09
48	lead-free	-1.13

3.2. *Discussion*

Table 3 shows how many category members are found among the high- and low-ranking groups (remind that the number of words in each category was 4). Once again, we have to be careful with comparisons given the low number of items, and we can only observe tendencies.

TABLE 3: Number of units of each category
in the low-inventiveness (A) and high-inventiveness (B) groups

category	A	B
simplex words (literal)		2
derivatives (literal)	2	
conversions	1	
back-formations	2	1
derivatives (with tropes)		2
compounds (literal)	3	
compounds (with tropes)		
simplex words (with trope)		
clippings	1	
blends		2
splinters		2
diverse		1

Obviously literal derivatives and literal compounds were judged to have a low degree of inventiveness (*lead-free*, which is highly transparent, is last in the ranking; *environmentalist* is in the low inventiveness group despite its

formal complexity). Here again, manufactured words (*nerd*, *klick*) were seen as inventive. Derivatives with tropes attracted attention, too (*commuterdom*, *poppers*); compounds with tropes, however, did not. Two of the blends (*glitterati* and *hacktivist* are the first two items in the ranking) and two of the units including splinters (*snowmaggedon* and *prequel*) were judged inventive.

4. General discussion

These two investigations are not without their limitations, mainly in terms of the numbers of stimulus units and informants. The number of words in each category, 4, in the second investigation did not overtax the informants' goodwill but it precluded the statistical testing of differences between categories. However, focusing the investigation on a smaller number of categories with more examples might have had the undesirable consequence of attracting the informants' attention to these categories, thus contaminating their judgments. Other limitations are inherent to the format of the questionnaires: what the subjects were given to judge were actually not the stimulus words in isolation but the stimulus-cum-definition items, and although care was taken to write neutral, dictionary-type definitions, one cannot exclude that some of the definitions had an influence on the responses, in particular by reducing the interpretive effort. Finally, judgments of inventiveness are metalinguistic tasks, which might not reflect exactly the informants' unconscious attitude to words (connotation, that is). Additionally, this attitude may be due to a certain extent to the knowledge of individual informants and their preferences for particular words. For instance, in the case of the derivative *poppers*, some subjects may be sensitive to the metonymy (the sound made by the container, or rather its opening, for the substance contained) – if they are aware of it, that is – while others may be influenced by the onomatopoeia present in the base and resulting in a lively, playful denomination.

These two exploratory investigations, however, have brought to light some tendencies that can be related to the hypothetical components of inventiveness that were suggested in the Introduction. Derivation and compounding are judged non-inventive when transparent, while blends and splinters, as well as manufactured words tend to be found inventive. This correlates with the word-formation / word-creation distinction. The distinction, however, has been criticized in its strict version on the grounds, for instance, that there is more regularity in word-creation devices than appears at first sight (blending is a case in point, see for instance Renner *et alii*, 2012), and also that intentionality and consciousness when coining a word cannot be measured (Haspelmath, 2002: 100). Bauer (2001: 71) thus notes that we have no way of clearly separating the

two types of lexicogenesis, and suggests that it is reasonable to consider word-formation and word-creation as prototypical categories. A prototypical view of the distinction is indeed supported by the present experimental results, with one possible exception: clippings, which are generally considered as resulting from a device belonging to word-creation (see for instance Booij, 2012: 21-22; Mattiello, 2013: 95), and none of which were judged highly inventive. It should also be noted that some conversions and back-formations which might have appeared inventive, at least to the (non-native-speaking) experimenter, like *to waitress* or *to bartend*, left informants indifferent. If such back-formations are deliberate coinings, this weakens the formation / creation distinction, but it does not do so if they result from unconscious reanalyses.

Another observation is that metaphor and metonymy contribute to inventiveness, as appears in derivatives and in the compounds in the first investigation. If, like Bauer (2001: 63-64), we consider that figurative extension should be part of creativity, this is another argument for a prototypical word-formation/ creation distinction, and inventiveness is the subjective facet of lexical creativity.

REFERENCES

Algeo, J. (1993): *Fifty Years Among the New Words: A Dictionary of Neologisms, 1941-1991*. Cambridge: Cambridge University Press.

Barnhart, R. K.; Steinmetz, S. & Barnhart, C. L. (1990): *Third Barnhart Dictionary of New English*. New York, NY: H.G. Wilson.

Bauer, L. (1983): *English Word-Formation*. Cambridge: Cambridge University Press.

Bauer, L. (2001): *Morphological Productivity*. Cambridge: Cambridge University Press.

Bauer, L. & Huddleston, R. (2002): "Lexical word-formation". In: Huddleston, R. & Pullum, G. K. (eds.), *The Cambridge Grammar of the English Language.* Cambridge: Cambridge University Press; 1621-1721.

Booij, G. (2012): *The Grammar of Words: An Introduction to Linguistic Morphology*. Oxford: Oxford University Press.

Concise Oxford English Dictionary, 12th ed. (2011): Stevenson, A. & Waite, M. (eds.). Oxford: Oxford University Press.

Gagné, C. L. & Spalding, T. L. (2009): "Constituent integration during the processing of compound words: does it involve the use of relational structures?". *Journal of Memory and Language* 60: 20-35.

Haspelmath, M. (2002): *Understanding Morphology*. London: Arnold.

Marchand, H. (1969): *The Categories and Types of Present-Day English Word-Formation*. Munich: Beck.

Mattiello, E. (2013): *Extra-Grammatical Morphology in English: Abbreviations, Blends, Reduplicatives, and Related Phenomena*. Berlin: de Gruyter.

OED = *Oxford English Dictionary*. <http://www.oed.com> - [last consulted 08/01/2012].

Renner, V.; Maniez, F. & Arnaud, P. J. L. (eds). (2012): *Cross-Disciplinary Perspectives on Lexical Blending*. Berlin: de Gruyter.

Ronneberger-Sibold, E. (2010): "Word creation: Definition, function, typology". In: Rainer, F.; Dressler, W. U.; Kastovsky, D. & Luschützky, H. C. (eds.), *Variation and Change in Morphology*. Amsterdam & Philadelphia, PI: John Benjamins; 201-216.

APPENDIX 1

Questionnaire (Investigation 1)

An anonymous questionnaire about word inventiveness

This questionnaire is part of an enquiry on inventiveness in vocabulary. The words below were randomly drawn from vocabulary which appeared mainly in the second half of the 20th century, and their definitions are given in case you are not familiar with some of them (some may be North American, others British only).

Please write next to each item a number from 0 to 3 corresponding to the degree of inventiveness you think played a role in the formation of the word. Use your spontaneous idea of inventiveness.

0 = not inventive at all
1 = not much inventiveness
2 = rather inventive
3 = highly inventive

If you are not a native speaker of English, please do not take part in this enquiry. Please do not ask other people what they think about the words. If you feel that your attention is drifting, it is all right for you to complete the task in several sessions.

trade off	=	a compromise between two desirable things that cannot be obtained together
crawler transporter	=	a large tracked platform used to move space rockets from the assembly hall to the launching pad
quark	=	a subatomic particle making up protons and neutrons
smoothie	=	a sweet, thick drink of fruit mixed with a dairy product
pollee	=	someone questioned in a poll
inflatable	=	that can be inflated
hominization	=	the process of evolution from prehistoric apes to modern man
shriek alarm	=	a portable personal alarm that emits a loud sound
monetarist	=	someone who adheres to the theory that money supply plays a key role in economic fluctuations

to pressure	=	to force by exerting psychological pressure
hang up	=	a psychological complex
fidelista	=	a partisan of Fidel Castro's politics
microsleep	=	a very brief involuntary sleeping period
do-it-yourself	=	building or repair work done at home by a non-professional
grassrooter	=	someone from the grassroots
ballute	=	an inflatable balloon-like device used like a parachute to airbrake bombs, etc.
moonlighter	=	someone who holds a secondary job in addition to his/her main one
light week	=	the distance covered by light in a week's time
egghead	=	an intellectual
maxicoat	=	a woman's coat reaching down to the ankles
extravehicular	=	that is or takes place outside a space vehicle
multihull	=	a boat with several parallel hulls
streaming	=	the grouping of students of similar ability
image-building	=	the creation of a positive perception by the public of an institution, company, etc.
black hole	=	a celestial body with such a huge mass that it traps light and is thus invisible
bullet train	=	a Japanese high-speed train
profit centre	=	a part of a company whose profits can be separately determined
spymaster	=	the head of a group of spies
systems analyst	=	someone who investigates computer requirements
glide bomb	=	a bomb that glides down to its target
condo	=	an apartment in a condominium (a collectively-owned building)
baby boom	=	a surge in the number of births
electronuclear	=	concerning a theory of physics that unifies the electromagnetic and the nuclear forces
hit man	=	a criminal hired to do harm to someone
ozone hole	=	the disruption in the Earth's ozone layer caused by certain chemicals.

helipad	=	a landing platform for helicopters
cryonics	=	the deep-cold conservation of human bodies in the hope that they can be resuscitated in the future
scam	=	a fraud
transaxle	=	a combined car transmission and differential
to amp (up)	=	to increase the sound level
to rotavate	=	to cultivate, using a rotavator (a cultivator)
terabit	=	1000 gigabits (a computing unit)

Appendix 2

Units tested in Investigation 2

- simplex words
  - *meme*
  - *prion*
  - *nerd*
  - *klick*
- simplex words with tropes
  - *virus* (computer)
  - *hub* (airport)
  - *drone* (plane)
  - *toxic* (debt)
- literal derivatives
  - *reader*
  - *containerization*
  - *environmentalist*
  - *intermodal*
- derivatives with tropes
  - *conspiracist*
  - *commuterdom*
  - *initialism*
  - *poppers*
- litteral compounds
  - *bird flu*
  - *diet pill*
  - *mountain bike*
  - *lead-free*
- compounds with tropes
  - *road map*
  - *footprint*
  - *centrefold*
  - *acidhead*
- conversions
  - *to sortie*
  - *to google*
  - *to page*
  - *to waitress*
- back-formations
  - *to curate*
  - *raunch*
  - *to bartend*
  - *to emote*
- clippings
  - *celeb*
  - *app*
  - *hood*
  - *nuke*
- blends
  - *hacktivist*
  - *infotainment*
  - *glitterati*
  - *webinar*
- splinters
  - *snowmaggedon*
  - *talkathon*
  - *prequel*
  - *petnapping*
- various
  - *fashionista*
  - *ultrabook*
  - *newbie*
  - *to badmouth*

APPENDIX 3

Questionnaire (Investigation 2)

NB The head text is identical to that in Investigation 1 and has been removed.

to badmouth	to speak ill of (someone)
raunch	energetic vulgarity
prequel	a film whose story predates that of another film in the same series
ultrabook	a very compact portable computer
footprint	the surface occupied by a machine, device, etc
intermodal	(adj) (of transport) that combines rail and road
environmentalist	a person who is an advocate of environmental protection
reader	a book-sized electronic device that displays the texts of books on a screen page by page
commuterdom	the suburbs
containerization	the introduction of containers in maritime and land transport
initialism	an abbreviation formed with the first letters of a sequence of words pronounced separately, like 'USA'
centrefold	a woman photographed in the nude for the folded central pages of a magazine
fashionista	a person highly interested in fashion or working in the fashion industry
to emote	to display emotion in an intense way
hood	a neighbourhood
acidhead	a heavy user of the drug LSD
snowmaggedon	a catastrophic general standstill caused by heavy snowfalls
diet pill	a pill containing substances supposed to cause weight loss
bird flu	a variety of influenza affecting birds
infotainment	a television program that mixes information and entertainment features
celeb	a famous person, esp. in the cinema or show business

to google	to search for information on the web, using the Google search engine
hacktivist	an activist who is also a computer hacker and breaks into systems to pursue his/her aims
klick	a kilometre
lead-free	(adj., esp. of petrol/gasoline) that does not contain lead
to waitress	to work as a waitress
road map	a set of instructions, esp. with a political goal
petnapping	the stealing of a pet for ransom
webinar	a seminar which takes place on the Web
app	a piece of software designed for mobile devices
glitterati	persons who are famous, rich and elegant
nuke	a nuclear weapon
conspiracist	a supporter of a conspiracy theory
hub	an airport used by an airline as the centre of its network
newbie	a person who is new in a job, on a team, etc.
meme	an element of behaviour, knowledge, etc., passed from generation to generation by imitation
talkathon	a long session of talks
poppers	a kind of illegal drug consisting of a liquid that must be inhaled
toxic	(adj.) (of debt) that involves a risk
nerd	a socially inept person who is obsessed by scholarly work
to curate	to act as the curator (of an exhibition)
drone	a remote-controlled airplane
mountain bike	a sturdy bicycle designed for use on rough and hilly tracks
to page	to call someone on a public address system or by walking round while displaying a sign with his/her name or calling his/her name
prion	an abnormal protein which is the infectious agent of Bovine Spongiform Encephalopathy (mad cow disease)
virus	destructive software that spreads from computer to computer
to bartend	to act as a bartender
to sortie	to make a sortie, i.e. to come out of one's defensive position, as in a siege