

A word in time:  
First findings from the investigation of dynamic text

*Antoinette Renouf*

University of Birmingham

## 1. Introduction

Increasingly, linguists are turning to corpus data to inform their descriptions of language use. Corpora of all shapes and sizes proliferate, each reflecting aspects of the past or present state of the language. Typically, these are 'static' in nature; that is, they are made up of texts selected from a given period of time, and treated as a homogeneous and synchronic entity. A recent development at Birmingham, however, is the investigation of text as a dynamic phenomenon.

The notion of dynamic text has a history (Sinclair, 1982; Clear, 1986) dating back to the early eighties, when Sinclair coined the term 'monitor corpus'. Building on that notion, I have devised a methodology, and set up a large, government-funded project in the Research and Development Unit for English Studies. In this, the focus is on a series of texts, viewed sequentially and chronologically. For years necessarily just a concept, this approach is now becoming a reality, largely through the availability of vast amounts of electronic text; and the scale of operation necessitates the automation of associated analytical procedures.

The dynamic approach to corpus study may not appeal immediately to the corpus linguist who is accustomed to working with a polished corpus of texts; dynamic data is often 'dirty' data, as Blackwell (1993) shows. Some linguists may also ask whether the effort of processing huge amounts of text only to discard them is justified.

Experience gained in the AVIATOR Project indicates that it is indeed worthwhile. In this paper, I will point to some of the linguistic facts that are emerging in early output from the first of AVIATOR's analytical tools, or software 'filters', known as 'Filter 1'. This is concerned with the identification of first occurrences at the level of the word.

For convenience, the sample data referred to here will be exclusively drawn from *Times* newspapers, within the period November 1990 to September 1991, although AVIATOR works in principle not just with journalistic text but with any electronically available textual data.

## Draft

### 2. Filter 1 findings: 'new' words

This filter automatically identifies 'new' words in a flow of data on the basis that they are not in a 'master word list' of 'existing' words. The filter automatically subdivides new items into proper nouns, abbreviations and acronyms, numerals and 'ordinary' words. Each batch of new items, once recorded, may be added to the master list, so that the system itself is dynamic.

#### 2.1 Rate of word coinage

The filter provides information about the degree of novelty in a particular text variety. For instance, it has identified 5,374 items as being new 'ordinary' word types in the approximately 2.5 million words of *Times* text published in March 1991.

In fact, the rate at which new words are found by our system is a reflection partly of the size and nature of the particular 'master list' used: and partly of the range and amount of text that is processed. With a master list that is still relatively small, the March crop is very large, but subsequent monthly batches of text will yield progressively fewer new words, as the master list is augmented; and gradually the pattern of new word occurrence will become clearer and more predictable.

#### 2.2 The commonest words coined in a given period

The filter carries out a cumulative word count, noting first and latest dates of occurrence. It shows that in the *Times* text from March 1991, 337 of the total 5,374 'ordinary' new types occurred more than once. Of these, the twelve most frequently-occurring items were as follows:

##### **videograb** 10

Mar-27: Daily Mail rushed in a high-tech **videograb** system for the Gulf war at a cost of **train-tick** 6

Mar-09: One of the pleasures is that a train-tick can never be verified.

##### **all-employee** 6

Mar-20: shares acquired under approved all-employee share schemes to be transferred self-**development** 5

Mar-07: receptive to the needs of its employees' training and self-development is one of the **train-birding** 4

Mar-09: a telescope for hours at a time. Train.blrdlog, done with the naked eye at 50mph, is **non-insane** 4

Mar-27: sleepwalking had been used as a self-evident illustration of non-Insane automatism.

##### **no-bailout** 4

Mar-04: powerful sanction would come into play, provided the no.ballout rule was strictly observed.

##### **videograbbing** 3

Mar-27: The Gulf war sped up **videograbbing**, but 1V companies are unhappy with the press

##### **top-shelf** 3

Mar-05: could be re-classified as a "**top-shelf**" publication unless it drops its advertisements

##### **semi-double** 3

Mar-13: are featured, such as 'Scentuous' with **semi-double**, white, pink-flushed flowers

##### **self-quotation** 3

Mar-26: Elgar's obsessively secret and personal **self-quotation** reaches out and moves the

##### **microflat** 3

Mar-21: modern units for construction, though the "micro-flat" is its first complete dwelling.

### 2.3 The categories of word formation found in a given period

The different types of word formation in the filtered output can be studied. Sections 2.3.1 - 2.3.8 below present a selection of items in text for February and March 1991 that reveals something of the range.

#### 2.3.1 Onomatopaeic items

Mar-10: open all four sash windows and **flumps** down on his cushion. "I get influenced by

Mar-07: rather than the Brothers Brooks. And he sings and **ho-hums** and chats to the Mar-17:

violence, with references to "a helicopter **whumphing** in the dunes", and the eyelets Mar-22: upper octaves. That voice, along with the **plonkily** rudimentary synthesiser backing

#### 2.3.2 Transcriptions of speech

Mar-24: an' that wis before thae bludy turtles. Ah'm tellin' ye, thae Tories are gonnae Mar-24: said about me. Britt was going, 'Ah, you look lurvely, darleeng'. How I survived all

#### 2.3.3 Puns

Perhaps as expected, puns are particularly apparent in headlines to articles, as in lines 2, 3, and 4 below:

1 Mar-08: national Symbolism to feature the sort of **boneless.wonder** figure drawing that

2 Mar-17: - Point not **myth-ed**; Pruffrock

3 Mar-19: **Cheque-ing** up; City Diary - Carol Leonard

4 Mar-21: Lamont awakes to find himself **transnormed**; Budget 1991 - By Graham

5 Mar-24: elected by proportional representation on a **last-past.the.post** system. "Once again

## Draft

### 2.3.4 Blends

Mar-28: would be on celluloid; the novel oozes **cinemaspirathons**. A series of overlapping Mar-06: after every barrage and has a habit of sending a **kissergram** by dipping the left shoul Mar-10: A raincoat. A briefcase. A tie. A **poserphone**. And pills. Well, if I had brought pills

### 2.3.5 Clippings

Mar-27: The Daily Mail rushed in a high-tech **videograb** system for the Gulf war at a cost of

### 2.3.6 Compounding

Mar-14: a humdinger of a writer who can twist the standard **schlock-horror** situation into a Mar-16: his theme, in his neatly crafted television news **sound-bites**. The night Sir Geoffrey Mar-17: transferring your brain, now known as **wetware**, as opposed to hardware and software

### 2.3.7 Derivation

#### 2.3.7.1 Alternative derivations

Some items that occur, particularly derivations, are aberrant, in the sense that alternative derived forms already exist, and these new versions do not appear to be justified in terms of doing a different job:

Feb-17: not know whether to cry or admire: his complete **indifferentness** and inattention  
Mar-14: spokesman said that people felt disfranchised because the results because the results  
Mar-17: most of the best items fall outside it, in a variety of **non-familiar** modes. Since all  
Mar-18: the Workington players tackled **mercifully**, and after a brief five-minute spell

It would appear that one productive rule has here overridden another.

#### 2.3.7.2 Humorous derivations

Other words, notably adjectival and adverbial, have presumably been, or were originally, coined to entertain the reader:

Feb-12: Remembered (Radio 5) were **actressly** modulated and suspiciously unteutonic.  
Feb-10: Frolicsome or **chucklesome** are the epitaphs that this puff of a novel  
Mar-05: not only fruitful, dearest," he replies, "but seedful, **flowerful**, and, yes, cormful, too"  
Mar-17: of all Jeffrey Archer, whose novels he still edits, **extracurricularly**. Will the popular  
Mar-17: In 1867 he renounced the academic vow of **bachelordom** and married one of them.

### 2.3.7.3 Suffixes

There are too many of these to enumerate, but by way of illustration of their creative use, sample instances of two common suffixes, ISM and DaM, for the period November 1990 to March 1991 are presented:

#### -ISM

Nov-26: Euro federalists to Thatcherite diehards. Thus **Majorism** enjoys a further unity  
Dec-02: The annual beanfeast of literary **tartanism** is always awash with Scottish sentiment  
Dec-04: parties to reduce their tendency to **cronyism** and extremism. The mechanism by  
Dec-09: Dundee 800 should be accused of Glasgow 1990 **bandwagonism**. Ward refutes any  
Dec-30: government nominally dedicated to a bit more **welfare-stateism**. As by 1992 it would  
Jan-04: (prejudice in favour of heterosexuals), **handicappism** or **disablism** (discrimination  
Jan-04: in favour of heterosexuals), **handicappism** or **disablism** (discrimination against the  
Jan-04: (discrimination against the disabled), **smokeism** (persecution of smokers), and  
Jan-12: porcine benefactor for the splendid thing it is. **Piggism**, irrational prejudice against  
Jan-19: clarity and range of the journalist with the **miniaturism** of the scholar that both  
Jan-20: setting), I found the relentless **miserabilism** through which it is filtered increasingly  
Jan-27: could be persuaded to end the deliberate **eco-terrorism**. This is very much expected  
Feb-02: of the features that bothers me is the bogus **peasantism**. The poem printed above is,  
Feb-13: reflex, are now leaden with archaic **far-leftism**. This is not without its black humour  
Feb-17: invented a new form of discrimination: **weightism**. ..TX.- Women over 11112 stone  
Feb-17: the lesson of other sports: that **shamateurism** does no good to the individual's  
Feb-26: - 26 February 1991 - A DOSE of **televangelism**, the American talent for bringing  
Mar-12: without intercourse. "This is the flipside of **abortionism**," Keith Davies, its campaign  
Mar-17: just follow anything, like sheep. ..TIC- **Sheepism**. At 1.15am at the Albert Hall,  
Mar-16: Fowler's eight pages on "that" and "**thatism**" explain why. His recommendation to

#### -DOM

Dec-02: As an unlikely bridge between **teeny-bopperdom** and thirty something CD-ism  
Dec-07: outward and visible badges of our **executivedom**. Old lady to yuppie self-importantly  
Dec-09: her original cabinet, through academe and **mediadom** and Foreign Office mandarins,  
Dec-09: it achieved the dizzy heights of Michelin **two-stardom**. The new chef, WiIli Elsener, is  
Dec-14: liberate ourselves from this stultifying, compulsory **coupledom**..TX.- ..XR.- Issue  
Dec-23: be more danger of him glamorising cheap **thiefdom**. We have also tried to cut down  
Jan-13: Centre and the broader spectrum of **dogdom** was on display. There was minimal bark  
Jan-20: across the country are slipping into **slumdom**, the bureaucrats will be housed in style.  
Jan-24: Jnr (played by Stallone Jnr) drifts into **punkdom** through lack of fatherly guidance.  
Jan-27: spons promoter, sUIVeying the field. ..TX.- **Chumdom**, it seems, would explain the  
Feb-07: too much, and punch too hard, for the sake of **best-sellerdom**. But some alarms  
Mar-02: late Seventies, tended towards the **clichédom** that previously afflicted curry-houses.  
Mar-31: Avenue who starve themselves into **chicdom** were lampooned as the Social X-rays

#### 2.3.7.4 Combining forms

There are again too many of these to enumerate. Some recurrent types from January 1991 to March 1991 are presented in Table 1.

Table 1: Recurrent types of combining forms

-RELATED	-FRIENDLY	-SPOTTING	-GRAM
cold-related	customer-friendly	plant-spotting	strip-a-gram
snow-related	dancer-friendly	advert -spotting	kiss-a-gram
dining-related	electorate-friendly	trend-spotting	strippergram
energy-related	family-friendly	grave-spotting	kissergram
rail-related	female-friendly	star-spotting	herogram
rental-related	people-friendly	choirboy-spotting	kissagram
savings-related	vine-friendly		fatladygramgram
size-related	women-friendly	-BUST.	strippingnunagram
tobacco-related	dog-friendly	boat-buster	Rambogram
topic-related	inflation-friendly	budget -busting	flanagram
turnover-related	media-friendly	gang-busters	fatladyagram
age-related		sanctions-busting	nun-strippergram
diet-related	-HOLIC.	spook-buster	
merit-related	ch0COholic	techno-busters	
mining-related	cruisaholics	bunker-busters	
pornography-related	sexaholics	bunker-busting	

### 2.3.7.5 Prefixes

There are many prefixes: some contextualized instances from March 1991 data for ECO- and EURO- are offered below:

#### ECO

Mar-01: The Road to Hell, and the grim warnings of eco-dlsaster and industrial decay have Mar-09: where do casually-uttered cliches, such as eco-terrorism, consumer-terrorism, and Mar-13: East Anglian style. There is also an "eco-pavillon" featuring a converted Victorian Mar-17: hardly keys with the fragile eco-system of Akamas. But if this is the cross-roads for Mar-18: Government would prefer to award "eco-labels" itself, with the EC as watchdog. Mar-31: CROFFING'S hour has come: it is eco-friendly and conservationist, a way of

#### EURO

In our sample data, ECO- tends to be hyphenated in more cases than EURO- does; see March 1991:

Mar-06: would be forced to redeem an ill-fated euroconvertble issue at a cost of Pounds Mar-19: for members of all the City exchanges, the eurobond market and corporate finance Mar-25: a rights issue. The two largest europderence holders, Lord Rothschild, the financier Mar-25: over the past two years, despite their appetite for eurosterllng issues, and might buy

### 2.3.8 Compounding

The filter system is set to regard hyphens as characters, so of the 5,374 new word types for March 91, over half are hyphenated. This gives some idea of the central role of compounding in the data, particularly since there are many more compounds that are not hyphenated.

The filter also frequently finds word-strings that contain more than one hyphen; the twenty longest for February and March 1991, in descending order, are as follows, beginning with a ten-word string:

- 1 Feb 07: a minimum-priced, quite-a-bargain-when-you-come-to-thnk-of-it double room is still a walk-on-
- 2 Feb-10: sling-and-fiddle-with-the-rope-walk-off affair
- 3 Jan 04: not a down-on-all-fours-pretending-to-be-a-dog dance, but a cross between a
- 4 Feb 08: It is the 'do-onto-others-as-was-done-to-me' philosophy.
- 5 Mar-14: The tale is spun out in effective will-they-won't-they-get-there-In-tlme fashion provokes complaints
- 6 Mar 08: of the can't-wrlte-songs-lke-they-used-to variety
- 7 Feb 17: I had this levada-walklng-as-a-doppelganger-of-skllng fantasy
- 8 Feb 10: story of her out-of-wedlock-pregnancy-and-parents-reaction to Today newspaper curiously from
- 9 Mar-05: windows as the red-rose-In-a-green-spray-lapelled party wafted by the whine, the snide aside, the
- 10 Jan-20: teeterlng-on-the-edge-of-a-consonant
- 11Mar-15: She has that you-gotta-laugh-or-else-you-cry tone in her voice.
- 12Feb-14: the abrupt ending smacks of author-racng-to-meet-the-deadllne
- 13Mar-31: His Essentialness is the born-agaln-and-agaln-and-agaln David Icke
- 14Jan-20: the Hemingway, experience-it-then-wrlte-it-down school of realist novelists
- 15Feb-17: for the Village Voice put it, there was "hanglng-from-the-ceillng-room-only"
- 16Feb-28: a very, very, tough mlnfield-with-barbed-wlre-fire-trenches type barrier
- 17Feb-26: it epitomises the nothlng-Is-too-bad-for-them public attitude to prisoners.
- 18Feb-17: A series of ludicrous, off-the-top-of-the.head suggestions
- 19 Jan-30: hotel is offering a special three-for-the-prlce-of-two bargain
- 20 Mar-16: think of the Inspector Morse stories as three-pounds-of-plck-'n'-mlx problems

These strings represent an important, but largely uninvestigated phenomenon. Their function seems to be to introduce an outside reference or allusion into the discourse, in a range of compressed grammatical guises, for stylistic effect. Syntactically, the strings often function as adjectival modifiers to a noun head. Several of the noun heads are abstract and three of the text-organising variety: see AFFAIR (line 2), FASHION (line 5) and VARIETY (line 6), which support strings which are clausal in structure. Some strings (lines 10, 12 and 15) function as nouns, but are again clausal.

### 2.4 Processes of word formation predominant in a given period

Two or more processes often combine in the formation of a word but, nevertheless, certain categories can be seen to feature strongly in our data: namely derivation and compounding.

### 2.4.1 Morphological features predominant within a category of word formation

Focussing on compounding, we can for instance discover which affixes and combining forms occur, and are most recurrent in a particular period. In March 1991, the combining form GATE, was not particularly productive; we find only the following four, original uses of the form:

Mar 11: reveal the true identity of Deep Throat, the Watergate informer, has been dismissed Mar 11: was, at best, a jaundiced view of Watergate. On the identity of Deep Throat, they Mar 17 Pulitzer prize-winning investigation into the Watergate scandal, which toppled Nixon Mar 18 interests during the difficult period of the Watergate crisis and Richard Nixon's

In late 1992, our filter would discover renewed press interest in and applied use of this form, with IRAQ(-)GATE, DIANAGATE, CAMILLAGATE and THRESHERGATE emerging in quick succession.

Among initial morphemes, UN (not U.N.) emerges as the commonest prefix. Hyphenated affixes and combining forms may also be identified, as is shown in Table 2.

Table 2: The 50 commonest initial forms in hyphenated word types for March 1991 (total 1,285 types)

non-	74	all-	14	ex-	12	multi-	9	not-	8
re-	42	post-	14	near-	11	five-	9	now-	8
over-	37	under-	13	out-	11	mid-	9	six-	8
self-	32	ever-	13	co-	11	sub-	9	double-	7
anti-	30	long-	13	hand-	10	cross-	8	fellow-	7
pre-	29	well-	13	high-	10	newly-	8	yellow-	7
half-	17	one-	13	pro-	10	white-	8	black-	7
mini-	17	three-	12	two-	10	best-	8	inter-	7
semi-	17	back-	12	counter-	9	book-	8	quasi-	7
four-	15	once-	12	single-	9	much-	8	small-	7

Table 2 shows that the most recurrent items predictably tend to be bound morphemes or closed set words, and rather short; becoming longer and more lexical as they become less frequent.

Table 3 shows the commonest final forms in hyphenated compounds. These common final elements are free morphemes, mainly of the 'combining-form' variety.

Table 3: The 50 commonest final forms in hyphenated compounds for March 1991 (total 1425 types)

-like	23	-making	8	-playing	6	-	5	-ups	4
-based	17	-led	8	-over	6	growing	5	-side	4
-style	15	-down	8	-only	6	-green	5	-round	4
-looking	15	-covered	8	-old	6	-general	5	-price	4
-sized	12	-up	7	-back	6	-faced	5	-point	4
-shaped	11	-loving	7	-run	5	-eyed	5	-play	4
-related	11	-in	7	-on	5	-clad	5	-paying	4
-free	10	-type	6	-off	5	-bound	5	-music	4
-makers	9	-set	6	-level	5	-writing	4	-minded	4
-out	8	-rich	6	-headed	5	-wise	4	-match	4
						-voiced	4		

### 2.4.2 Lexico-grammatical trends

Within the general categories of word formation, we can see which lexico-grammatical trends are predominant. Within derivation, for example, it is possible to observe the general directions of word-class shift. We note that nouns derived from nouns feature significantly in our data. In March 1991, we find, among others, the following:

Mar-24: for superannuated governesses is written in fluent brochures). finally (my account -iUs

Mar-10: But I would prefer any surrender to European bureaucratitis to occur under Labour -ography

Mar-10: Furia is not, in fact, into much biography or even careerography, but a lyricist's life -dom

Mar-02: late Seventies, tended towards the clichedom that previously afflicted cuny-houses. -ene

Mar-31: publisher, Dr Frands, and Thatcher not the ex.leaderene but ascertain David.

-ery

Mar-24: Ex-con playwright Duncan Roy boasted that his infamous fraudery, posing as Lord -isation

Mar-23: destroyed the basis for a unified Iraqi state. Lebanonisation seems the likely, but -en

Mar-31 many of these young blue bloods will be top notchen in time and Julie singles out

## 3. Conclusion

This paper has given a brief overview of our early work, concerned with monitoring the first occurrences, or reappearances, of words in text. Just from these initial findings, it will be clear that there is enormous potential in our automated approach. Our first filter throws hitherto inaccessible

Draft

light on traditional accounts of word formation in English, and will surely lead to a new description in the coming period.

DRAFT