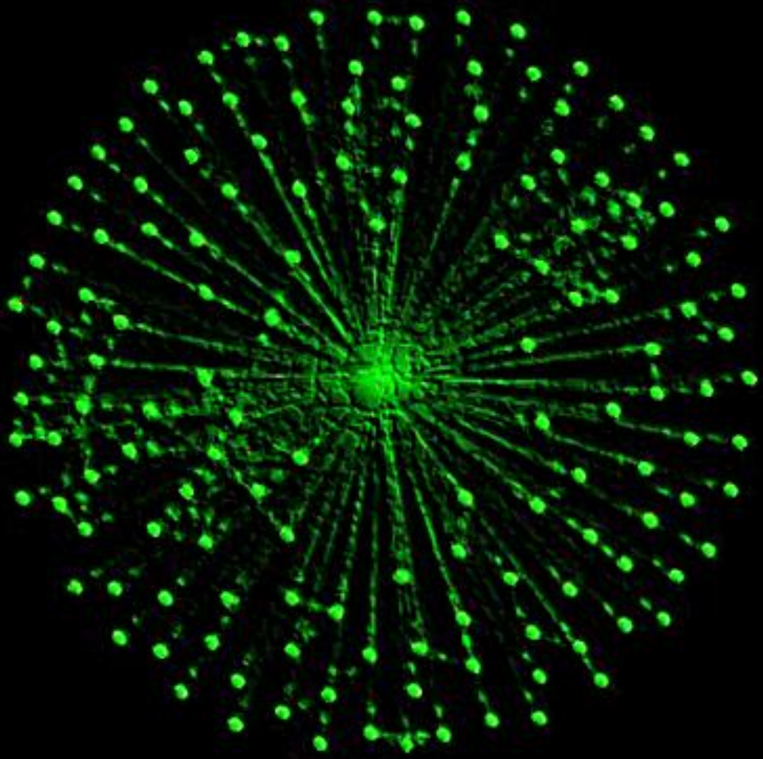
PALESTINE IN RUINS AND IN MASS MEDIA



ON METHOD

MANUEL ECHEVERRÍA

2021-11-21

On Method is the second paper of *Palestine in Ruins and in Mass Media*. It consists of an empirical and a theoretical part: First section reviews the fixed-point search engine introduced in the first paper of the series (Echeverría, 2021). In conclusion, algorithms employed so far yield suitable signatures of the narrative of a given topic. Comparison with refinements are implemented with a much larger set, of 733 articles on the Israel-Palestine conflict, from the biggest Swedish media outlets. These refinements deal with networks of words. The sample covers May 2021 when violence escalates. In summary, both methods and results from the first paper are robust but the refinements are more reliable.

The second section provides a mathematical formulation of the framework. A general setting is considered where a probability measure, for the robustness of a given ranking, is derived. This scheme deals with *configurations of unknown complexity*. In summary, naïve, basic or mechanical rules with plenty of flaws can be successfully used to gain relevant information from language, good enough for hypothesis testing. Minimal prior knowledge about grammar or language is required. Moreover, the scheme allows for arbitrary or inconsistent rules. In the end a probability measure reflecting the usefulness of the method is given regardless. The solutions to these problems and the scope of the framework justify a general theoretical discussion on society, and provides probabilistic accounting perspectives.

KEYWORDS:

Applied Mathematics; Theory of Society*;*
Algorithms; Media Analysis; Accounting

INTRODUCTION

This paper is a methodological piece in a series about media and ideology. The first part analyses journalism on the Israeli occupation of Palestine when violence escalates. It departs from a dataset on Swedish nation-wide outlets; both commercial and public service; in print and online. The analysis is mainly carried out with algorithms which are scrutinised, and subsequently given mathematical formulation in the second.

Information Production Theory postulates potential for diversity in news coverage and opinions, among commercial newspapers, in low-stakes topics. This series investigates the extent the Israel-Palestine situation allows for journalistic heterogeneity in states which have a distinctively independent, neutral or low-stakes stance towards an ongoing conflict.

Although Sweden officially has a relatively safe position towards the conflict in relation to the international community, the first part of the series on Palestine showed that Swedish Television (SVT) followed the state line closely.

Minister of Foreign Affairs made numerous ambiguous statements in the calibrated language of diplomacy but was clear on four points. Firstly, the state acknowledges the illegality of settlements and evictions in view of international law. Secondly, the conflict is not understood in terms of legitimate resistance to hostile Israeli occupation. Thirdly, the state acknowledges violence on both sides, but depicts Palestinian as indiscriminate acts of terror, whereas Israeli violence is seen as legitimate responses against specific enemy targets. Finally, insistence on precise Israeli violence results in cognitive dissonance, or justification by blaming the victims.

SVT propagated all these talking points with remarkable vigour and detail but deviated in one conspicuous aspect. There was a deviation on the last aforementioned key topic, but towards the Swedish conservative/right-wing spectrum. Association of Palestinian victims with terror went beyond the stance of the state. In conflict with wide-spread claims about left-wing bias, the position of SVT resides in the Swedish far right in this regard. These findings suggest that the overarching topic is more high-stakes, less neutral or more dependent than the safe official position reveals. Theoretical deliberations are postponed to forthcoming papers.

DATA

The main sample of the escalation in May 2021 is from the six biggest nation-wide newspapers. Four in print, and the two most successful online editions.

T1   Main Dataset: Six Nation-Wide 1-31 May

Outlet	Official Orientation	#Items*	Edition	#Alexa**
Dagens Nyheter	Independent Liberal	114 (126)	PRINT	
Svenska Dagbladet	Uncommitted conservative	90 (101)	PRINT	
Göteborgs-Posten	Liberal	63 (67)	PRINT	
Sydsvenskan	Independent Liberal	77 (81)	PRINT	
Aftonbladet	Social Democrat	173 (173)	WEB	2742(7)
Expressen	Uncommitted Liberal	129 (185)	WEB	11836(26)

* Noise removed. **Initial** dataset in parenthesis, my estimates.

** Overall ranking, 90 days prior 2021-07-31, domestic in parenthesis.

All of these are considered right wing by Swedish standards, except Aftonbladet, commonly considered centre-left. The average time spent on the sites is 3:04 and 9:16 respectively.

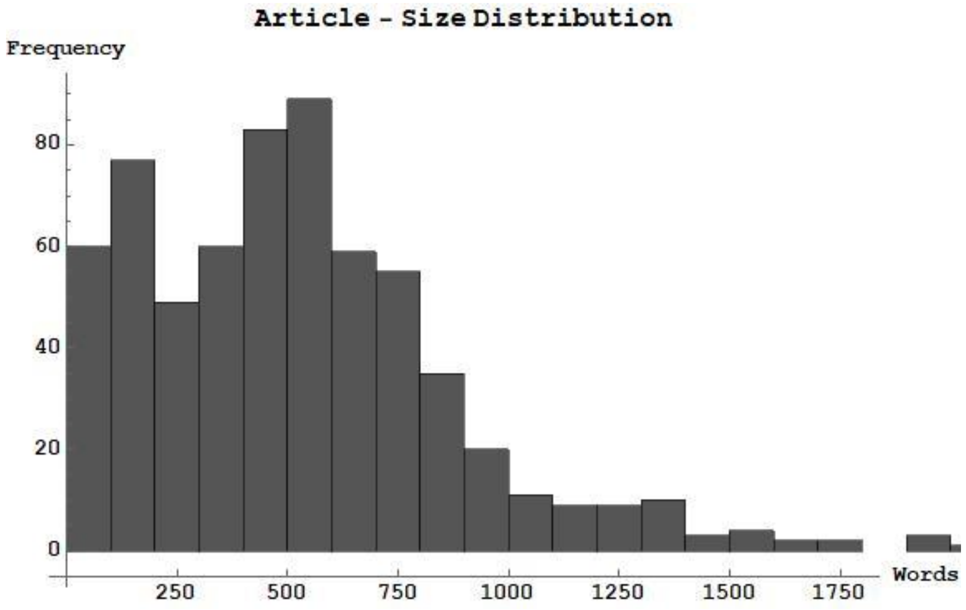
The initial main dataset consists of 733 articles May 1-31. The initial dataset was obtained through Retriever Research online database with searches on Palestine, Israel, corresponding citizens, and inflections. After cleaning data from blanks mostly announcing other items in the set, and omitting one irrelevant outlier of five standard deviations, 646 articles remain in the operational set. The median number of articles per outlet is 102, standard deviation 36. Only texts are analysed in this study. Including titles, the average article is 537 words, median 491.5, standard deviation 400. *Departures from the Retriever Research figures* is due to the superior accuracy of the methods in this paper[1]. This is also true for statistics on the initial set.

T2 Main Dataset: #Words/Item

Outlet	Average	Median	SDEV
Dagens Nyheter	608	518	539
Svenska Dagbladet	515	502	358
Göteborgs-Posten	546	561	297
Sydsvenskan	481	479	329
Aftonbladet	569	486	433
Expressen	474	438	302
Total	**537**	**491.5**	**400**

The sizeable spread is affected by the coexistence of shorter news items alongside opinion pieces and essays. The dataset has been processed to exclude metadata and noise associated with register procedures from the source research database. Titles are included to keep the information as close as possible to the original publications and perform relevant content analysis.

[1] E.g.: A new subset was downloaded with *exactly the same keywords*. Still, it had **less** items than it should when compared with a set from a month earlier; duplicates are a minor issue; and so is erroneous counts on nr of items.



POLITICS, WAR & CONFLICT

This study also makes use of Retriever Research's categories, *Politics* and *War & Conflict* which serves as a suitable point of reference for relevant opinion pieces and news items. The union of these consists of 276 (331) items; with items of 442 words on average, median 412, and standard deviation 310. Multimedia does not crowd out text size for Aftonbladet in this sample (T4).

T3 Politics, War & Conflict: #Items

Outlet	Official Orientation	#Items*	Edition
Dagens Nyheter	Independent Liberal	35 (39)	PRINT
Svenska Dagbladet	Uncommitted conservative	51 (62)	PRINT
Göteborgs-Posten	Liberal	25 (26)	PRINT
Sydsvenskan	Independent Liberal	22 (24)	PRINT
Aftonbladet	Social Democrat	48 (49)	WEB
Expressen	Uncommitted Liberal	39 (73)	WEB
SVT	Officially Independent	56 (58)	WEB

***** Noise removed. **Initial** dataset in parenthesis, my estimates.

One very important point regarding online datasets is warranted. Careful comparative analysis with Swedish-Television data from the first paper of this series (T8) reveals that the 29 items in the archived dataset at Retriever Research, differ substantially from the most visible versions at the time collected in the foreign news page – also 29. Only half of the first 12 coincide. This goes beyond updates as completely different items have been recorded. This means journalists to some extent can publish one set of accessible articles as history unfolds, and then archive a different set for future historical record.

T4 Politics, War & Conflict: #Words/Item

Outlet	Average	Median	SDEV
Dagens Nyheter	543	498	407
Svenska Dagbladet	414	451	283
Göteborgs-Posten	613	622	257
Sydsvenskan	567	523.5	269
Aftonbladet	537	474	351
Expressen	387	365	243
SVT	237	246	149
Total	442	412	310

HYPOTHESES

H1: Divergence from the state line exceeds pre-emptive openness of less than 5 % measured in Echeverría (2018; 2020).

H2: Heterogeneity exceeds Swedish Television along ideological orientation as measured in Echeverría (2021).

H3: Homogeneity within journals, heterogeneity across outlets with different ideological orientation.

H1 is the main theoretically founded hypothesis. All conceivable outcomes are theoretically and politically important. H2 is more interesting for political reasons, but useful for further theoretical discussions due to the close match between the official state line and Swedish Television. H3 is formulated to find out more about the specific modes of labour-market discipline and opinion management. It is informative of how hiring and assignment strategies are arranged. E.g. diversification of opinions within a journal as a way to compete for audiences. Specific rhetorical techniques will be discussed in view of Echeverría (2018; 2021).

Please note that from a statistical point of view, it does not matter if the hypothesis is rejected, not rejected or even accepted, as long as the procedure to do so is correct. Only H1 has been formulated in line with the general thesis in this series, as it is derived from rigorous theory. The rest may or may not correspond to my personal beliefs of the state of affairs.

METHOD

If one would treat the tempting habit to always report support for hypotheses at stake as a rhetorical convention, there would be no need to condemn pretence in science in this regard. Ultimately, as long as the method is transparent or correct, it is potentially informative, and useful for further inferences, regardless if it is ad hoc or not. In the end, either a study is replicable or not. Either the hypotheses are derived from theory by means of valid deduction – testable in independent studies – or not.

Unnecessary energy is still spent on condemnation and lengthy treatments of this theoretical triviality, as if one spurious study could forever obstruct all scientific efforts of a field by exhausting and spoiling all data for ever. The habit of discarding theories based on superficial methodological grounds is a game which belongs in the realm of politics, not reason. Perhaps it is just a yardstick of the prevalent intellectual culture that so many seem to have revealed preferences for dumb, deaf and blind scientists incapable of seeing patterns in the world around before formally theorising – over sharp intellects, constantly forming theories from an acute perception of even the slightest trends in everyday flows of data – for obscure philosophical reasons. Certainly, this remark exposes the absurdity of mistaking calculated political trickery, for honest indignation in the pursuit of truth. This argument is not necessarily sensitive to unique datasets. Observe it would be irrational or blatantly dishonest to play a game of reason by raising aforementioned objections in domains where my response appears to be inadequate.

Thus the need to spend a minimum of effort to settle this antiquated methodological issue, and associated petty objections, in a pair of paragraphs. Especially in view of the ridiculous political friction, and vicious attempts to suppress this otherwise rather conventional media study, on the bloody Israeli occupation of Palestine.

Quantitative and qualitative analysis of text is computed with specialised software developed by me. Commercial software is employed for simulations, statistical inference, and to test results or code.

The first part of this paper presents descriptive statistics resulting from an exploration of journalism by means of algorithms; most notably a fixed-point search engine. These algorithms run through vast amount of data related to the main threads, guided by theory.

Mathematical Articulation of the Framework connects the methods to a general theoretical discussion and makes clear that statistical reasoning is at the core of the algorithms. However, the complete-study nature of these automated inquiries relies less on statistical inference to estimate parameters for now.

RESULTS

The descriptive statistics in this section begin with a fixed-point search engine which tracks words most closely associated with a *trace*, consisting of an initial set of keywords, sentence by sentence. This procedure is referred to as a *hunt* or *tracking* in what follows. The terms directly associated with a trace are called *direct association.*

Thereafter, the words most commonly related to the *direct association* within the sentences are tracked. The resulting set of words is the *equilibrium association*, which typically approaches overall frequencies of most used words (See Echeverría, 2021a).

The searches or *Hunts* with the fixed-point engine were found to give suitable signatures, succinctly capturing the main topics of a given discussion, when evaluated with quantitative and qualitative content analysis (Echeverría, 2021). This time, the algorithm is examined with a list of alternatives in order to evaluate its general ability: In particular ones departing from paragraphs, groups of similar words, or both.

The associations are presented with their respective shares of the top five in parenthesis. The main set (647): Terror and associated inflections occur 264 times in 165 unique items, 0.4 times/item on average. The *direct association* (DA) to *terror*, for this benchmark algorithm is:

Hamas(**41%**)–Israel(**25%**)–Gaza(**12%**)–EU(**11%**)–Terror branded(**11%**).

The corresponding *equilibrium association* is:

Israel(**36%**)–Gaza(**22%**)-Hamas (**19%**)–Israeli(**14%**)–Israel’s (**9%**).

Table 5 compares this benchmark algorithm with refinements, able to account for groups of similar words like inflections and composite words. The results suggest that the basic algorithm gives a remarkably accurate account of the rankings when compared to more advanced procedures in the main dataset. For the latter, each word in the heading is the leader in a network of similar words (subjects). For instance, *terror* is merely one word with 8 counts, and *terror branded* is an independent word with a count of 44 in top 5 in the benchmark. In the refinements, introduced in this paper, *terror* is the leader of a network with a count of 264 and 56 subjects, *terror branded* among these.

A random sample of 10 translated words from the group with the leader *terror is*: *terror organisation*, *terror act*, *terror actions*, *terrorism*, *IS-terrorist*, *terror tunnels*, *terror groups*, *terror movements*, *terror branded*, *terrorise*.

T5 Direct Associations by Algorithm Specification

Top 5 BC|NC*	**TERROR****	**ISRAEL**	**HAMAS**	**GAZA**	**ORGANISATION**	**TOTAL**
#Words Top 5	264	200	178	80	60	782
Real Prob. NC(a)	36%	27%	24%	11%	8%	742
Virtual NC (b)	34%	26%	23%	10%	8%	782
#Words Top 2-6	**ISRAEL**	**HAMAS**	**GAZA**	**ORGANISATION**	**ISLAM**	**TOTAL**
Virtual NC (c)	35%	31%	14%	10%	10%	573
#Words Top 2-6	**ISRAEL**	**HAMAS**	**GAZA**	**ISLAM**	**EU**	
Intuitive NC (d)	35%	31%	14%	10%	10%	567
#Words Top 5 (BC)	**HAMAS**	**ISRAEL**	**GAZA**	**TERROR****	**EU**	
Basic Algorithm	41%	25%	12%	11%	11%	408

* **BC** = Basic/Benchmark Algorithm: *whole words*. **NC** = Network Algorithm: *word groups*.
**Terror is a group for refinements, but the word *Terror-Branded for benchmark cat.*
(a) Probability(A or E) must subtract 5% intersection of *Terror* & *Organisation* (40)
(b) Double count allowed. Shares of total Virtual - Thus proper 'double count' avoided
(c) Intersection of just 1. Thus almost coincides completely with real probabilities.
(d) *Organisation* dissolved (weaker bargaining position): 2/3 of its counts also in *Terror*. Former drops below top 5 when these are removed, while *Terror* prevails.

All algorithms in the table track sentences. T5 gives an account of top 5 and top 2-6 because the benchmark algorithm departs from the same encompassing 264-hit terror-keyword, but only seeks whole words thereafter. Thus, it usually gives info on 1+5 top terms.

Virtual shares are in important regards more real than shares of a total without double counts, of course*. Real Probabilities* can be thought of as proper shares obeying the appropriate concept of probability. See Appendix (A1) for details.

A parsimonious way of grasping the robustness of a *given ranking* or *order* is naturally achieved by counting the number of groups it takes to beat the lowest member of the order, i.e. its weakest link. In a non-stochastic setting, the convention can e.g. be to start with those outsiders with greatest counts adjacent to the weakest link, and check how many it takes.

A measure of robustness of an order towards outsider constructs of *unknown complexity* is given by:

The probability a sample, of randomly selected groups of outsiders, has a combined count greater or equal than the weakest link.

This probability is informative about the remaining combinations, potentially merged in ways, unaccounted by the simplest grouping and ranking rules of an algorithm. The number of merged groups it takes to challenge status quo, i.e. be on par or beat the weakest link, is the *breakthrough number*. A count equal or above is a *breakthrough;* the successful coalition is the *challenger*.

In other words, the probability of a challenger with breakthrough number (of at most) B, given information about the outsiders, is the generalization of the natural idea. The probability of a challenger with breakthrough number seven is 5.0 % without further information[2].

[2] Computations show that knowledge about a handful adjacent outsiders will give anticipated but minor alterations in probability. See *Mathematical Articulation of the Framework*.

Table 6 compares the distribution of frequencies for three informative words, and their inflections in commercial media, with Swedish Television (May 7-16) in the first paper of this series.

T6 Relative Frequencies Part I & II

	TERROR	PALESTINE	OCCUPATION	TOTAL
PAPER I (a)	52%	31%	17%	29
SHARE II (b)	27%	54%	19%	981
SHARE II (c)	23%	55%	21%	433

(a) SVT Paper I, May 7-16 **(b)** May 1-31 (c) May 7-16

Clearly, the habit of using the term *Palestine* is much greater in commercial media, in relation to the other two. Moreover, there is a slightly higher propensity to use the word *occupation*. Due to the accuracy of the database, this latter discrepancy is also to be considered as a reliable result.

Table 7 displays the ratio of the number of articles containing *Occupation* over the number of articles containing *Terror*.

T7 #Items with Keyword [Occupation/Terror]

	PAPER I	RATIO II (a)	RATIO II (b)
Occupation/Terror	50%	64%	82%

(a) RATIO II: MAY 1-31 **(b)** RATIO II: MAY 7-16

The tendency in table 6 is also carried over to the number of unique articles mentioning occupation or terror respectively. Interestingly enough, *occupation* is mentioned even more frequently, in relation to *terror*, in the first half of the period in the commercial newspapers.

POLITICS, WAR & CONFLICT (PWC)

In the *Politics, War & Conflict* subset – 279 items after processing – *terror* occurs 144 times, in 93 unique items, about 0.5 times/item on average. The overall direct association has most weight on terror-branded Hamas:

Hamas(**40%**)–Israel(**23%**)–Terror *branded*(**13%**)–Gaza(**13%**)–EU(**12%**).

The corresponding equilibrium association May 1-31 is:

Israel(**35%**)–Gaza(**21%**)–Hamas(**20%**)–Israeli (**14%**)–Israel's(**10%**).

Table 8 shows the direct association to compare with paper 1 May 7-16. The similarity is striking. However, on the set of all outlets, *terror* (*branded*) is just outside top 5 of the basic algorithm, due to a greater variety of inflections employed in the sample. The two first rows are based on the dataset of Politics, War & Conflict *including Swedish television* (SVT).

T8 DA by Algorithm Specification

Algorithm BC|NC*	Politics, War & Conflict | Top Distribution May 7-16
NC All	Terror(**28%**)–Israel(**24%**)–Hamas(**21%**)–Gaza(**12%**)–EU(**7%**)-Islam(**7%**)
BC All	Hamas(**37%**)–Israel(**26%**)–Gaza Strip(**14%**)–EU(**12%**)–Rockets(**12%**)
NC SVT2	Terror(**29%**)–Hamas(**24%**)–Israel(**21%**)– Rockets (**13%**)–Several(**13%**)
BC SVT2	Hamas(**29%**)–Terror(**29%**)–Rockets(**16%**)–Several(**13%**)–Organisation(**13%**)
BC SVT1	Hamas(**32%**)–Terror(**32%**)–Israel's(**12%**)–Gaza(**12%**)–Rockets(**12%**)

***BC** = Basic/Benchmark Algorithm; **NC** = Refined/Network Algorithm. SVT1&2 are results for state television in paper 1 and this paper respectively. *Terror* is as usual a group for refinements, but the word *Terror Branded* in basic cats.

Figures are presented in compact molecular form but computed as in T5. Note that the *database* for Swedish-Television in paper 1 (SVT1) and Retriever Research (SVT2) differ significantly. The former is more representative of the most visible versions of the articles at the time.

BREAK POST BOMBING AP & AL JAZEERA?

The responsible thing to do is to check if there is a break before and after the May 15 Israeli bombing of the offices of Associated Press and Al Jazeera. Therefore the first and second half of May are contrasted. Commercial and state television are treated separately within the Politics, War & Conflict dataset in order to improve commensurability. *In summary*:

Direct-Association Signature: When SVT is removed, the overall narrative on terror is similar, but the direct associations indicate commercial newspapers tell a slightly different story. This is especially true for the basic algorithm which displays a progression from a more particular account of the state of affairs, to a more general in the second half. This is to a much lesser extent true for the network algorithm, but there is a slightly more undivided attention towards religion and the European Union in commensurable periods.

T9 Israeli Bombing of Press Offices

Algorithm BC|NC*	Politics, War & Conflict | Top Distribution
NC 1-31	Terror(**33%**)–Israel(**25%**)–Hamas(**22%**)–Gaza(**12%**)–Islam(**9%**)
BC 1-31	Hamas(**39%**)–Israel(**25%**)– EU(**14%**)- Gaza(**11%**)–USA(**10%**)
NC 1-15	Terror(**30%**)–Israel(**26%**)–Hamas(**20%**)–Gaza(**14%**)–EU(**9%**)
BC 1-15	Hamas(**31%**)–Israel(**27%**)–Gaza Strip(**17%**)– EU(**14%**)– Rockets(**10%**)
NC 16-31	Terror(**36%**)–Israel(**23%**)–Hamas(**23%**)–Gaza(**9%**)–Organisation(**9%**)
BC 16-31	Hamas(**41%**)– Israel(**18%**)–Gaza(**15%**)–USA(**13%**)–EU(**12%**)

***** **BC** = Basic/Benchmark Algorithm; **NC** = Refined/Network Algorithm (Virtual) as in Table 8.

PARAGRAPHS

A new dataset with *intact paragraph structure*, in *Politics, War & Conflict*, was downloaded about one month after the first. Although downloaded with *exactly the same keywords*, the raw file on *commercial outlets* (183 after processing) nevertheless contains 58 items less than it should. The processed files differ with 37. Thus, this set is still suitable for comparing the algorithm but somewhat less reliable in terms of external validity.

T10 Summary Statistics 1st & 2nd Download #Words/Item

Politics, War & Conflict – Commercial Outlets

Dataset	Average	Median	SDEV
PWCC I (220)	494	474	318
PWCC II (183)	514	495	313

As is evident form T11, the algorithms tracking words in paragraphs instead of sentences yield essentially the same results. This is especially the case for the network-based algorithm.

T11 Basic, Network and Paragraph Algorithm

Algorithm: BC|PC|PN*	Politics, War & Conflict II | Top Distribution
Basic (BC) 1-31	Hamas(**40%**)–Israel(**21%**)–EU(**17%**)–Gaza(**12%**)–USA(**10%**)
Paragraph (PC) 1-31	Hamas(**39%**)–Israel(**25%**)– Gaza(**14%**)- EU(**11%**)–Israeli(**10%**)
Network (PN) 1-31	Israel(**29%**)–Terror(**26%**)–Hamas(**23%**)–Gaza(**13%**)–Rocket(**9%**)
Sentence (NC)	Terror(**37%**)–Israel(**22%**)–Hamas(**20%**)–Gaza(**12%**)–EU(**9%**)

BC = Basic Algorithm: Whole words. PC = BC with Counts on paragraphs. PN = Network version of PC. NC = Sentence-based Network algorithm, as described in T5, for reference.

CHECKING ALGORITHM INTERPRETATION

An additional dataset of equal size was generated to conduct tests. In this dataset, both algorithms are much closer in terms of composition, suggesting a development of the narrative from particular to more generic concerns involving bigger players and regions. In particular, the network algorithm suggests the narrative in the first half is dictated by an antagonism between *Terror-branded Hamas* that *rules Gaza* and *Israel*. (A2 in Appendix)

The following is conjectured: Commercial outlets focus more on international reactions instead of Hamas rockets, *especially in the second half*. This reflects a progression from a more specific to a more general narrative. *Swedish television maintains focus on rockets* but with international reactions. **Individual Searches:** The trends remain similar but with relative less weight on occupation. SVT increases its relative weight on *terror*.

RESULTS

As expected, the interpretations closely correspond to content – both when the particular sentences counts are based on were checked – and in more comprehensive random samples of articles. The logical intuition is that when a sample has been conditioned with a search on a relevant trace, relevant terms tend to be attracted. The most objectionable conjectures are about the connection between *Hamas*, *terror* and *rule*, these are therefore checked.

Sentences: Qualitative content analysis reveals that all instances of Hamas are directly connected to *terror*. Thus, the coexistence of the two terms is not a coincidence – Hamas is presented as synonymous with a 'terror organisation'. The term *rule* was as expected connected to Hamas, the *rulers* of Gaza. Turns out that this conjunction exists in 90 % of the occurrences of the word *rule* and similar inflections.

Random Sample on whole Articles: The same thing is true when a more comprehensive analysis based on a random sample of 20 items in the first half was carried out. This time, 11 were related to *rule* proper, and 90% (10/11) to Hamas as above. This shows that the same interpretation is valid when whole articles are checked at random.

More generic narrative. It was conjectured that the second half of the month had a more general discussion with international actors. There is truth in that conjecture when relevant sentences were checked. However, the international actors are invoked to underscore the terrorist status of Hamas. USA and EU are more frequently cited, in the second half of the month, but as authorities that have branded Hamas as a terrorist organisation.

In conclusion, the perhaps most natural interpretations that can be made at first glance of the molecular composition corresponds to the actual content of the associated sentences of the tracking procedure. More comprehensive random samples of articles yield similar results. However, it is important to bear in mind that a carefully chosen initial trace will focus the results on relevant topics, which will not necessarily hold true on the aggregate. This result is notwithstanding reflected in the discrepancy between direct association and equilibrium association.

SUMMARY

Comparisons of the basic fixed-point engine with refinements show it is a suitable summary statistic, remarkably close to more intricate alternatives throughout the study. The main difference is that the refinements rank words in groups of a wide range of inflections and composite words. Thus, there is a methodological continuity between the first paper and this one in spite of the changes.

There are starker differences in journalistic output. Succinct descriptive statistics generated by the algorithms are so far consistent with the general hypothesis of a more diverse commercial press, when comparing commensurable periods. However, this will be settled in forthcoming work.

The individual searches on *Terror*, *Palestine*, and *Occupation* made in the first paper also differ in the same direction. The relative weights between the two first are switched, and the overall distribution between the three is more balanced. *Palestine* is now much more frequently used than *Terror*, in turn now much closer to the frequency of *Occupation*.

Contrasting the first and second half of the month adds to the impression of a more independent journalism compared to SVT, or a lag between state narrative and diffusion in commercial outlets. The difference in emphasis on terror increases in the second half of the period. The commercial outlets have less immediate connection to the initial agenda-setting statements of Minister of Foreign Affairs in this particular regard.

This continuity has traces of self-similarity in terms of most frequently used words. Especially at the level of sentences and paragraphs, but perhaps less so at the level of whole articles and aggregate journalistic output. Exact degree of self-similarity, and the underlying behavioural dynamics generating it, is a topic worthy of more attention in the future.

Earlier discovery and analysis of exact *propaganda fractals* in journalism (Echeverría 2018, pp.166-174) showed that these were directly related to repetition of falsehoods and dogmatism. In view of indications of more diverse journalism, other causes to self-similarity in topic distribution must also be taken into consideration if significant.

It is my sincere intention not to insult the reader with the following advisory: Labels naming theory so far and in what follows, do not necessarily correspond to the everyday use of the words, e.g. *autarky* below.

MATHMATICAL ARTICULATION OF THE FRAMEWORK

This paper considers operations on information structured as a language. The approach in this paper deals with the following conditions or problems:

(1) **Convoluted or idiosyncratic information.** Expressions may be created with inconsistent rules.

(2) **Limited knowledge & autarky.** Operations must work without complete information about the language, preferably without additional information to the dataset under study. (no complete dictionary or grammar-book condition)

(3) **Transparency requirement.** The algorithms employed must allow for low-cost scrutiny, and preferably be simple or tractable.

This setting opens up for serious challenges, for example: Convoluted info and limited knowledge (1 & 2) allow for expressions of *unknown complexity* which nevertheless must be accounted for. Transparency and autarky (2 & 3) limits the scope of at least a subset of machine-learning procedures which have been trained on large datasets exogenous to the dataset to be researched (2), or may yield good results but with virtually intractable rules (3). Low-cost scrutiny may e.g. be defined as an amount of effort/information processing to achieve transparency, below a certain share of what is required to compute the operation on the dataset.

This setting is theoretically motivated in terms of the comprehensive scope (1), or with regard to common-knowledge requisites for scientific inquiry (2 or 3). There are also much more mundane reasons, for example: State oppression and hacking may perturb incentives and the marginal propensity to be online to acquire databases. Hardware or other limitations may make it more efficient to program a tailor-made solution than to incur an economic or human-capital investment in alternative technology. In other words, this setting corresponds to an outsider or underprivileged person’s situation.

PROBABILISTIC APPROACH

The operations considered here are related to the frequency distribution of expressions, and ranking of these accordingly. Depending on convention the scheme can be described in two or three steps.

- **(i)** Form groups of expressions
- **(ii)** Rank groups formed in **(i)**
- **(iii)** Compute robustness of rank formed in **(ii)**

Once the first two steps have been computed, the groups simply have been ranked in accordance with the sum total of the frequencies of each word in the group. Consistent with requirements (1-3), only a horizon of a couple of groups at the top are properly checked in step (i & ii), the rest are just the results of a simple algorithm. Such a list will therefore contain information about the distribution of counts by groups, and potentially more elaborated *processed* information about the top. The problem is that although the top (or order) may be satisfactory, partially by invoking arbitrary rules in addition to the simplest algorithm, there is no way of knowing if the simplest of rules have ordered the rest intelligibly, and there are no reasons to conclude such state of the world is the case for sure (1-3).

Nevertheless, the distribution of groups makes it possible to account for network-configurations *of unknown complexity* by counting the number of groups it takes to beat the lowest member at the top, i.e. the weakest link of the order. A generalisation of this natural idea is:

The probability a sample, of randomly selected groups of outsiders, has a combined count greater or equal than the weakest link.

A configuration of groups not in the order (*outsiders*), which is greater than the weakest link is called a *challenger*. The size of the random sample is denoted B, and is called the breakthrough number.

The ranking can be understood as a list containing frequencies of group sizes. The latter are in the set {1, 2, …, r_{max}}, where r_{max} is the *heaviest* group of the order in terms of counts – not number of subjects in the groups. Therefore, the probability of finding a challenger as a result of a random sample, of B drawings without replacement, is Multivariate-Hypergeometrically distributed (MVH). The probability of a challenger with breakthrough number B, with a weakest link of count W is denoted:

$$\textbf{F1} \qquad \mathbf{P(xr >} \mathrm{W} \mathbf{\,|\, I)},\ \mathbf{x} \in \mathbf{MVH}(\mathrm{B}, \ddot{\mathbf{g}})$$

Where $\ddot{\mathbf{g}}$ is a vector with the distribution of *group sizes* in terms *of counts* – i.e. the frequencies of distinct group sizes in the sample; **x** is the vector of groups, **r** is the vector with corresponding counts for the groups in **x**. The dot product **xr** is required to be such that each group variable is multiplied with its corresponding count. Furthermore, **I** is conditional info. E.g. a logical circuit over information on incompatibility of groups in the vicinity of the order, retrieved from the horizon in the previous steps. Such condition could lower the probability of a challenger. Useful information can be extracted from the graph-matrix containing vectors of connections for each word. The length of such truncated or directed vectors, informs about the degree distribution of connections (see next section). Extreme outliers can be used to rule out too large groups *in terms of subjects*, which together with the information from the top, narrows down or maps out the domain for challengers.

In practice a 5-standard-deviation threshold identifies expressions which always have been ruled out so far, usually mistakes or with three or less characters. In conclusion, this approach and the formula pins down what is to be regarded as relevant *exogenous* information about language in this setting, but much can be deduced or at least plausibly conjectured from the dataset without additional info. Although this remark also underscores that F1 is based on a useful model, not 'reality', this point cannot be emphasised enough.

GROUP FORMATION

From (2) there is at most an incomplete dictionary and grammar. The sample consist of about 700 items, roughly 390k words and 2.25 million characters in total. To speed up research, a list of less than 600 words, 60% between three and six characters, with about 2,7k characters in total was used to delete noise. This list contains connectives, the alphabet, digits etc. and is more than enough to produce the results in the first part. If such list is feasible, and there is no theoretical interest in the viability of full autarky, skip this section and go to *The Graph of Connections*.

Exogenous information in aforementioned list can be drastically reduced with a partial model of what meaningful information or a language is. But in the end, even if this model is flawed, the general scheme of the previous section computes a summary probability measure for the reliability regardless. This is true by first principles. Hopefully the reader will be able to immediately think of tens of exceptions to these simple rules as he or she goes along, and will be able to write down at least a hundred or so within minutes. As all working hypotheses, the actual usefulness must be checked empirically. The results in the first part of this paper show that a few simple rules are remarkably useful.

Information is considered as generated by a set of atoms. A sequence of *atoms* is information. In particular: *ordered* subsequences of atoms are delimited by markers; these units are called *molecules*, including composites. Sequences of molecules make up *substance*. Substance is generated by a configuration of rules or markers, possibly with unknown or inconsistent rules.

The point of departure is a dataset to be researched, e.g. the one explored in this paper. This set is for purpose of exposition regarded as representative. If the sample was complete, then there would be a full *dictionary*. When all the rules or conventions necessary to form substance are known, then there is a full *grammar*. Molecules identified in the sample are facts, and (2) holds.

There is a basic assumption: Facts are likely to be useful to express something and can therefore be used to inquire without knowing their exact meaning. The operations to find out about *implied grammar* on this set is by and large restricted to intersections and complements *between molecules*. The *ordered* sequence of atoms is of interest. If subsequences between two molecules coincide, i.e. they intersect in terms of an ordered sequence of common atoms, then they have a *common denominator*.

A related conjecture is that similar molecules belong together in terms of expressions. In other words, the conjectured existence of variations of meaningful molecules generated by the underlying rules. Variations or *inflections* are (to some extent) identified by the atomic complement of two molecules. In these cases, the common denominator is a *fact* – but the atomic complement is not. Conjectures can be made based on efficiency requirements or statistical inference. E.g. the complement is over a certain fraction of the factual molecule. In particular if it is heavier/longer than the factual molecule it would tend to be less likely as an *attachment* to denote inflection.

If the complement is factual molecules or factual molecules with inferred *attachments*, it is a composite. If the grammar is somewhat consistent, the set of attachments in the inflection complement will tend to be recurrent, and if efficiency is assumed also limited. Under these circumstances, the set of inflection complement corresponds to a residual or *rest* of attachments.

If a complement degenerates such that it is neither a fact proper, inflection, or a fact with attachment, then it is *strange*, i.e. rubbish.

In summary, complements are employed to pin down viable related groups of molecules by checking whether intersecting expressions are composites, inflections or strange. This process of joining and deletion is like a Tetris game with a probability of meaningful or unrelated connections, where the former is conjectured to be greater than the latter for valid groups.

THE GRAPH OF CONNECTIONS

The graph of connections between words with common denominators is central and will be derived in this section. Connections in the graph are made up by common denominators. To get an idea of its structure and purpose, note that lighter/shorter molecules will tend to have plenty of common denominators/connections to other molecules. Moreover, for a sizeable share of their connections, the intersections will fully cover them (complete). For now, a complete common denominator is denoted by one for the smallest, and those between zero and one are truncated to zero. When this information is stored in a matrix, it results in a directed graph of connections.

More precisely: For purpose of exposition, consider molecules as ordered sets, or tuples of atoms **<**a_1, a_2, …, a_n**>**. This can also be seen as a set with indexed elements, where the index or coordinates denote position of the elements proper. An overlap **of threshold n** is an operation O_{ij}^{n} over ordered sequences of atoms which belong to molecule m_i, $m_j \in$ **Ω** respectively; where **Ω** is the set of all molecules. In a more general setting, n_i is a threshold value for the size of an intersection between any ordered subsequence of a molecule i, with another. The index of n will be supressed lest it is needed.

For now, it is enough to pay attention to the case where n is the number of atoms of m_i to be selected from the first element of m_i up to the atom with index n. This sequence, O_{i}^{n}, is then compared in n steps, element by element, to ordered sequences in m_j. If the elements proper are not equal for each of the n steps, then an overlap is not returned (necessary condition). For purpose of exposition, consider the special case when the operation starts from the elements with lowest index and takes steps in increasing direction, and only the first overlap is returned, which is denoted O_{ij}^{n}. (*****)

Consider the corresponding weights $w_{ij}=\frac{|O_{ij}^{n}|}{|m_i|}$ and $w_{ji}=\frac{|O_{ji}^{n}|}{|m_j|}$, where $|m_k|$ is the total number of elements in molecule k, or largest index. Note that $|O_{ij}^{n}| \equiv n$ if overlap exists, 0 otherwise. In general, $|O_{ij}^{n}|$ & $|O_{ji}^{n}|$ could both be positive and differ in various ways depending on n, direction etc. In the special case (*****), with $n_k=|m_k|$ a heavier molecule is not *overlapped* by a smaller. Therefore:

$$w_{ij}=\begin{cases} \mathbf{1} \text{ if } O_{ij}^{n} \neq \emptyset \\ \mathbf{0} \text{ otherwise} \end{cases}$$

A connection is formed when a molecule is smaller and overlaps, in which case its weight has the value one, and zero otherwise. In more general cases, w_{ij} potentially is at least a triple with information on whether the intersection is above threshold or not (1 or 0), and the relative size of the molecules in an intermediate case, e.g. when w_j is smaller. This also is informative about the size of the overlap. The connections for each molecule to others are collected in vectors $\mathbf{c}_i$. The matrix **M** of such row vectors is the graph of the network of connection. The number of positive elements in $\mathbf{c}_i$ is the number of connections of m_i which can be used to spot statistical anomalies given the empirical distribution. Note however that when the set is restricted to a given topic, the keyword corresponding to that topic will tend to also be an anomaly.

Basic groups are formed on the *set* **F** of facts, which clearly usually yields a much smaller number of molecules to handle than the *list* of facts. The row vectors in **M** correspond to candidate groups with subjects connected to a group leader, when mapped to the set of molecule tuples. The set of atomic complements between the leader and the subjects C_{ij}^{n} is the corresponding operation to O_{ij}^{n} with the same specifications as above but returns the remainder of m_j when the intersection is subtracted. If the atomic complement is rubbish, it is thrown away (see previous section for definition). In practice, groups with anomalous lightweight leaders with many connections but few independent counts, are usually dissolved.

This deliberately simplistic procedure yields remarkably sensible groups. The ranking is made by counting each unique molecule in **F**, in the dataset of interest, with an algorithm which counts networks of words corresponding to the row vectors above. This is what most word-processors do when they highlight your searches on a document. Only a horizon of an order of size *o* plus a number of potential outsiders adjacent to the order is treated, until the order is stable for this horizon, with the aforementioned basic-group procedure. This ranking can be *reduced* by removing inflections from the horizon, but I leave the details on this to the reader.

MINOR REMARKS ON THEORY AND PRACTICE

This is not a paper on algorithm efficiency, computability or fixed-point semi-decidability. In any case, the algorithm under scrutiny is proved to be computable by construction and concerns a decidable set, when restricting attention to the first step of direct association. Thus, details on specific rules such as to determine what happens when rankings tie etc., are left out. However, such exercises are not excluded in future footnotes if there are serious points to make. This is a paper about empirical findings and a general theoretical discussion departing from scheme (i)-(iii).

Step (iii): Notwithstanding the special case of uniform distribution, it stands to reason that sizeable empirical distributions may distinguish between most counts at the top even if they are close. Hence, there will still be almost as many options to draw from even when information about a few major ones is given. Therefore, the conditional probability may not differ much in such cases. Indeed, computations show that minor perturbations in likelihood of a challenger is the case in this paper. Truncation parallel to Scott, Sturges or Wand binning procedures could be used transparently if deemed necessary. Moreover, approximations, such as the Zipf distribution, are convenient.

Step (ii): This part can of course be related and substantially improved by consulting the literature for algorithms on networks. **Step (i):** The same holds true for grouping and the literature on contract theory or network formation which suits theoretical or empirical work best.

Although there are analogies between method, theory and empirical application of theory, there is a caveat: what is methodologically trivial may be regarded as a serious theoretical or empirical problem in other domains. **For example:** Empirical work can make use of all information available where theory may assume limited information. The framework in this paper makes it possible to list all group members and sizes, while specific algorithms make use of only some information. However, interesting theoretical considerations may assume an even narrower information set or explain partial use of information as imperfect recall or resource-constrained effort exertion.

Empirical remark: Relatives to the geometric distribution are evident at all levels, from word-frequencies in the networks of molecules, to aggregate journalistic output. These are well-known language and network properties.

GENERAL REMARKS

There is a rather obvious intuition associated with social science and other domains but there is more to it than parables. A more subtle point is the parsimonious conceptualisation of general conditions shaping institutions:

a) Multiplicity of rules, and distribution of knowledge about theses b) how information interacts with what is to be or not to be regarded as subjective d) how to deal with these characteristics when rules also are inconsistent.

Note that this framework not only allows complicated atomic constructs both at the level of molecules and substance. Complexity is also allowed in terms of rules of inference or grammar, and these claims have been substantiated.

APPENDIX

(A)COUNTING CONVENTIONS

Probabilistic perspectives on accounting are given througout this paper in line with Echeverría (2020). This section promotes transparency with a rationale and a straightforward example for the counting conventions. In particular to virtual or double counts, real counts and probabilities in T5. First row of A1 gives the basic counts of each category, A&B shows that category A and B have two elements in common.

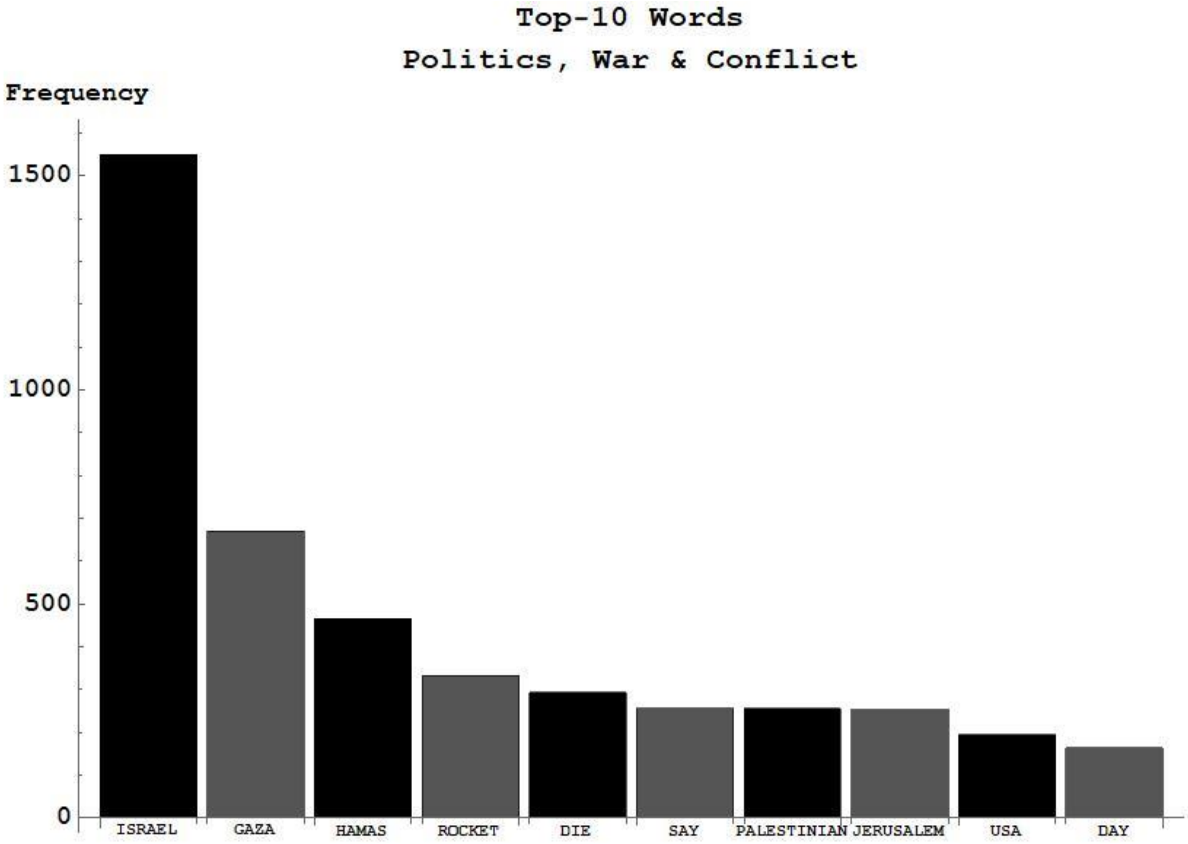
A1 Double Counts

	A	B	C	A&B	REAL TOTAL	DOUBLE COUNT
COUNTS	4	4	12	2	18	20
VIRTUAL SHARE	0,2	0,2	0,6	0,1	1	
REAL ‘SHARE’	0,222	0,222	0,667	0,111	1,111	
	A	**B**	**C**	**A&B**	**AVBVC**	
PROB	0,222	0,222	0,667	0,111	1	
	AVB	AVC	BVC	A|B		
	0,333	0,889	0,889	0,5		

Although some would intuitively prefer a *Real Count,* which avoids double counts, over *Virtual Count,* which is based on double counts, the latter is on all accounts more natural than the former. The resason is that simple substraction of the conjunction will not do. It makes no sense because the groups cannot both not contain the intersection, and it is potentially arbitrary if one does but not the other, and erroneous if both do because it would then be double count. As the table shows, shares or a probability measure based on substracted total (18) is degenerate. In contrast, virtual shares do not alter the underlying numbers and the shares or probability measures have the usual interpretations because the doubles are considered as different (virtual) entities. **A proability concept to accounting is the correct approach** as employed in the row ‘PROB’ if *Virtual Counts* are disregarded for some reason.

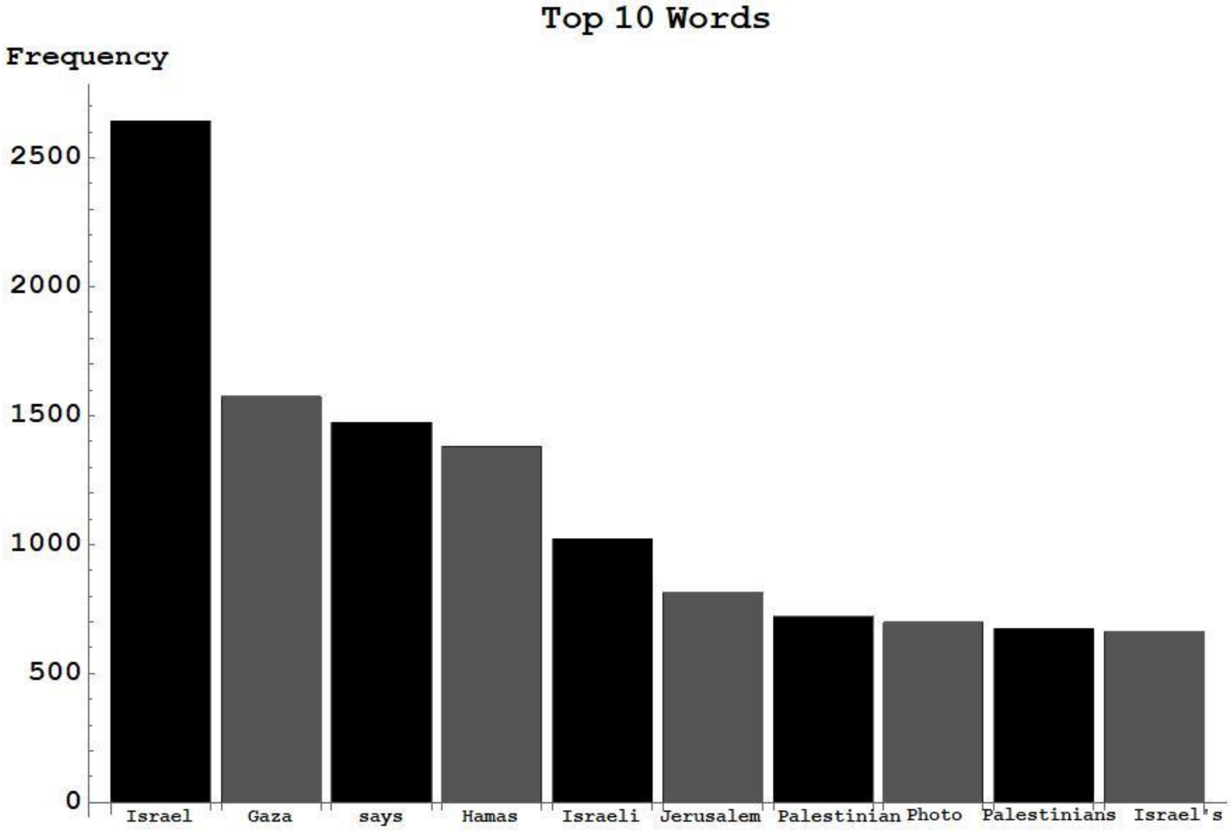
BASIC & REFINED RANKING

For reference, the following graph on the most frequently used words in the subset *Politics, War & Conflict*, was created from a random sample of articles, half the size of the full set (n = 140).



It gives a fair picture of the actual distribution, and the usefulness of the basic grouping algorithm, which is an improvement but has limitations. There is one exception to the rule. *Palestinian* and Palestine and related inflections should arguably be in a group on second place. However as their inflections do not have a common (empirical) denominator, they are treated as different groups.

The chart above is much more informative in terms of content compared to one based on whole words as below. Observe that the tracking algorithms are much closer to each other, with or without refinements.



A2. Test of Algorithms

A2 displays the molecular structure based on an alternative dataset generated for an earlier version of this paper, published 2021-10-20.

A2 Israeli Bombing of Press Offices

Algorithm BC|NC*	Politics, War & Conflict | Top Distribution
NC 1-31	Terror(**31%**)–Israel(**26%**)–Hamas(**22%**)–Gaza(**13%**)–EU(**7%**)
BC 1-31	Hamas(**39%**)–Israel(**24%**)– Eu(**13%**)– Gaza(**12%**)–Terror Branded(**12%**)
NC 1-15	Terror(**28%**)–Israel(**28%**)–Hamas(**21%**)–Gaza(**14%**)–Rule(**9%**)
BC 1-15	Hamas(**32%**)–Israel(**27%**)–Gaza Strip(**17%**)–Rockets(**12%**)–Terror Branded(**12%**)
NC 16-31	Terror(**33%**)–Israel(**24%**)–Hamas(**23%**)–Gaza(**11%**)–USA(**9%**)
BC 16-31	Hamas(**41%**)– Israel(**18%**)–Gaza(**15%**)–USA(**13%**)–EU(**12%**)

* **BC** = Basic/Benchmark Algorithm; **NC** = Refined/Network Algorithm (Virtual) as in Table 8.

REFERENCES

Echeverría, M. (2018). *WikiLeaks' Unforgivable Liberalism*. The Internet Archive. Libertarian Books – Sweden. Stockholm/Bergamo

archive.org/details/WikiLeaksUnforgivableLiberalism

This is an online-friendly (2.3mb Baskerville/Caslon). The official version was published by Libertarian Books – Sweden (**87.0 MB, MINION**).

Echeverría, M. (2020). Footnotes on the Figures of Suppression: Footnote 1 – Propaganda Accounting.

Echeverría, M. (2021). Palestine in Ruins and Mass Media in *Information Production Theory*. Forthcomming.

First Version Published:
2021-10-20