Flaming: A White Paper

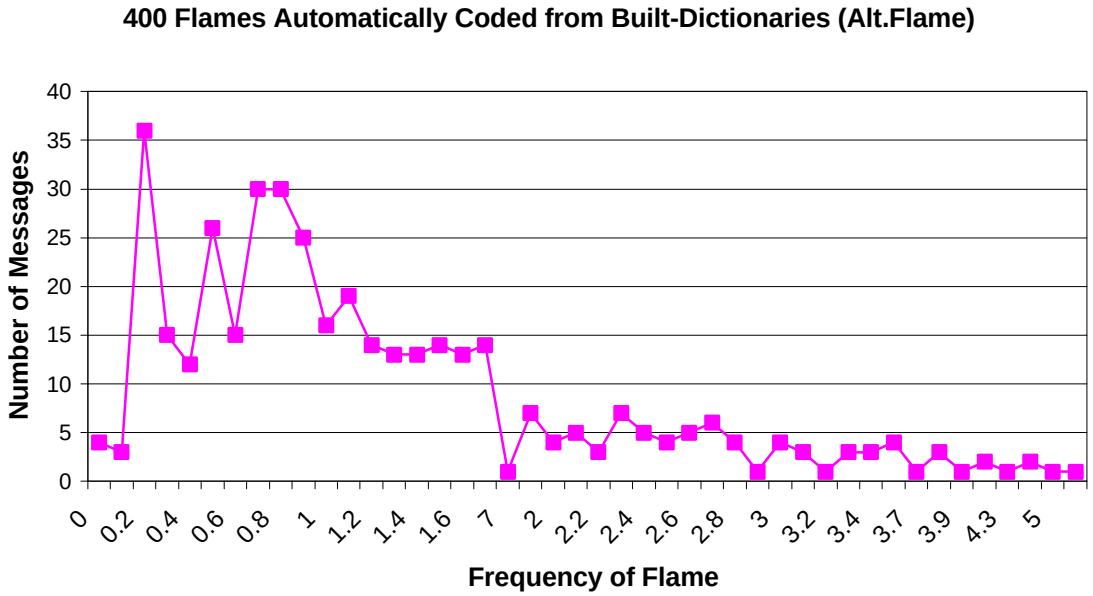
David Kaufer
Carnegie Mellon, English

Flaming: A White Paper

David Kaufer
Carnegie Mellon
Department of English
June 2000

1 Flaming: Introduction

This paper reports on a systematic study of flaming behavior, drawn from examining and coding some 1,200 messages from the newsgroup alt.flame. Alt.flame is a newsgroup where flames are composed and publicly broadcast. While I have no data on the readers of this newsgroup, the participants are people who flame, as opposed to people who study or analyze flaming. The newsgroup thus gave me a direct look at the practice of flaming. At the same time, it gave me a look *only* at public rather than private flaming. Unlike private flaming, public flamers know their flames are being read and digested by an anonymous readership. This condition may put pressure on flamers to make their flames artful and witty as well as inflammatory. It very likely makes their flames longer. Before I began coding flames, I assumed that flaming "hot spots" in a message would involve the majority of the message. After building a dictionary of some 3,000 phrasal flames from a sample of 400 messages, I tested the dictionary on a new sample of 400 un-coded messages. This re-testing allowed me to expand and refine the dictionaries. It further allowed me to study the density of flames per message. The breakdown of flame density in the messages of alt.flame is presented in the chart below.



As the above chart indicates, in alt.flames, flaming patterns are not terribly frequent occurrences in the context of the entire messages composed. The vast majority of flames in my sample had less than 1.7 flaming words per 100 words. Private rants against an individual might be expected to be shorter, less literate (i.e., "not surrounded by conventional, non-flame, discourse"), and exhibiting a higher frequency of flames (as a function of their shorter length) than public flames. However, that remains only a conjecture at this point and not systematically explored in my research.

2 Flaming Defined

For the purposes of my research, I define flaming as computer-mediated communication designed to intimidate the interlocutor by withholding the expected courtesies of polite communication. Sometimes the withholding of respect takes the form of direct aggressiveness against the interlocutor. Often, it takes the form of gross insensitivity and bad taste, not only against the interlocutor but also against the culture at large. The expression of hatred through bared teeth seems in and of itself to have a frightening and intimidating effect on human beings. Flamers seem capable of intimidating solely by expressing their hatreds, even if the listener meant to be intimidated is not the personal target.

3 Flaming vs. the Perception of Being Flamed

There are two implications of this definition. First, flaming is somewhat in the eye of the beholder and the beholder's perceptions of the speaker and the context. In a context where harsh language is expected (a locker room), the presence of extreme vulgarity and lewdness is by itself not felt as a flame. To be felt as a flame, a message must come across as an effort to corner, isolate, humiliate, and, in general, intimidate the interlocutor by stripping her of dignity. The mere occurrence of language that overlaps with intimidating or offensive behavior is insufficient to account for flaming. There must also be a contextual motive to hurt or offend and the reader must perceive the intent for what it is.

Second, the electronic medium both seems to promote flaming while also walking the fine line between flames and games of pretend-flaming among friends. These conditions are probably intertwined because the physical stakes (of violence) are less consequential in electronic interaction than in face-to-face physical interaction. This may make it more appealing to flame on-line. Yet the lack of a known shared context in electronic communication makes the speaker's base intentions (to intimidate) less clear. Does the writer mean to be intimidating or simply falling into crudeness with no underlying motive to offend or intimidate another human being? We can generally discriminate these conditions in the context of face-to-face interaction, where the speaker's face and body are available along with the words. It is harder in email or on newsgroups, when we can read only the writer's words.

There is invariably a probabilistic component to most flames. That is to say, their occurrence leads to the suspicion of a flame but not the certainty. Certainty, or near certainty, depends upon knowing the full context and the full context involves knowing how the perceiver of the flame relates to the flame.

4 Flaming and Context-Sensitivity

In private flaming, the impact of the flame (viz., intimidation) is often felt in the utter absence of any language that is conventionally associated with flaming. Imagine an employee who is responsible for getting a report out that is late in delivery. The employee's supervisor knows it is late, knows the employee already feels bad it is late and, in a fit of anger, sends the following abrupt e-mail: "When is the report coming

out?" The language, on the surface, is an innocent question about the future. Yet the employee feels it as rubbing salt on his wounds, a direct challenge to his fulfillment of his work assignments and, perhaps, his fitness as an employee. The message is meant and deeply felt as a flame, though it bears none of the characteristic linguistic patterns of a flame. No computer program can possibly capture such flaming. Yet the most personally hurtful flames are probably of this type.

Because of these implications, one can never be sure whether actual flaming behavior is accurately captured solely in the language conventionally associated with flames. Indeed, a devious flamer may issue a flame, intending to intimidate, with the ever-present deniability clause that he or she did not think the reader would find the message intimidating or offensive.

Since a computer program can't know the context of an alleged flame, the best a program can do in the context of uncertainty is to remain focused on the conventional language of flaming and to approximate the effect of context.

1. occurrences of highly offensive language

2. frequency counts of language that falls into a pattern of flaming

In cases like 1, the language pattern "F… you Bastard!" can be associated with a flame merely by the weight of this single occurrence. In cases like 2, the occurrence of a single oath or curse word ("shit", "screw it") would not register flaming behavior. But if there are enough of these words, or the words constitute a certain percentage of the whole text, the suspicion of flaming can be highly enough activated to be flagged.

To accommodate both kinds of cases, I chose to divide every flame dictionary into two parts: Regular flame dictionaries and what I call High-flame (hereafter H) dictionaries. Regular dictionaries contain phrases the phrase parser detects through frequency. A single occurrence of such phrases won't be sufficient to register as a flame. The flames will need to appear as a percentage (e.g. .05%, 1%, 2% etc.) of the entire text. The longer the text, the more occurrences of flaming patterns will be needed to impact the flame meter. The limitation of this approach is that for very short or long texts, the frequencies of flame patterns may be skewed very high or very low respectively. By way of contrast, The H-Dictionaries contain phrases that the parser detects as "high" solely on the strength of a single occurrence. Unavoidably, these are subjective judgments, though nothing prohibits testing on user groups. The optimal calculus for associating flaming phrases with a metric of offense (such as chili-peppers) should be developed in interaction with user communities.

5 A Flame Coding Hierarchy

I have divided flames into a coding hierarchy. The coding hierarchy makes it easy to maintain the phrasal dictionaries without losing the structure and consistency of the overall coding scheme. At the top level of the hierarchy, flames are classified into categories. Categories divide into phrasal dictionaries. Categories and the phrasal dictionaries they organize are now overviewed.

5.1 Categories

Flamers can focus on one of three entities: the flamer's own hard-hearted beliefs, the shortcomings of the reader, or the shortcomings of groups within the culture. Each focus represents a different category and a different weapon of intimidation against the reader of the flame.

5.1.1 My Pit-Bull Brain!

These phrases intimidate through the flamer's self-regarding behavior. By emphasizing the hardness of their own convictions, flamers can discourage and even "freeze" a reader from wanting to talk back.

5.1.2 Screw You!

These phrases intimidate through direct aggressiveness against the reader.

5.1.3 Screw Everyone!

These phrases intimidate by attacking subgroups in the general culture. This is the basis of racism, sexism, homophobic language and cultural slurs. Readers can be intimidated by the speaker's callousness and zest for degrading other human beings.

5.2 Phrasal Types

Categories differentiate different phrasal types. Each phrasal type is divided into regular and H-dictionaries.

5.2.1 My Pit-Bull Brain! Phrases

The dictionaries in this category capture language that carries the impression that the speaker is close-minded and uninterested in reasonable dialogue, dialogue where views can be freely explored, exchanged and elaborated. The language rather suggests a pit-bull, with glazed eyes, spewing the products of a closed and ugly mind.

5.2.1.1 Opinionated Comments

These phrases consist largely of noun phrases depicting the speaker's close-mindedness. An example in the regular dictionary is "retarded idea;" in the H-dictionary, "f***ing bullshit."

5.2.1.2 Opinionated Acts

These phrases consist largely of verb phrases registering hard-headedness with a trace of action-orientation. An example in the regular dictionary is "screwed up;" in the H-dictionaries, "f***ed up."

5.2.1.3 Heated Denial

These phrases undercut statements asserted previously, typically by the reader him or herself. Usually with a "not." An example in the regular dictionary is "I'm not about to. . ." In the H-dictionary, "that's bullshit."

5.2.1.4 Paranoid

These phrases indicate that the speaker sees him or herself cut off from a hostile and conniving world. An example in the regular dictionary is "why is everyone. . ." In the H-dictionary, "they're all f***ing. . ."

5.2.2 Screw You! Phrases

These phrases indicate a speaker on the aggressive against the reader of the message. The speaker is not just hard-headed and hard-hearted but on the attack. The second person "you" is consistently represented or implied in all the phrases across these dictionaries.

5.2.2.1 Face Threats

These phrases put the speaker in the face of the reader as part of a direct challenge. An example in the regular dictionary is "I am sick and tired of your. . ." In the H-dictionary, "f*** you."

5.2.2.2 Incapacities

These phrases resemble face threats, but cite the reader's shortcomings in the context of the attack. There are a characteristic set of incapacities flamers reference when they wish to refer to previous senders. Flamers, for example, often question the compositional abilities of previous flamers. So a commonly cited incapacity is "had to forge" (her message). Flamers also call one another "IQ challenged" and "cranially challenged." I have not, at this writing, created an H-dictionary for incapacities.

5.2.2.3 Taunting

These phrases bait the reader in addition to condemning or ridiculing him. This often occurs with a question intonation to provoke another turn from the reader. Examples are "then how come?" and "do you?" with many variations of punctuation (e.g. do you???>>>). There is currently no H-dictionary for taunts.

5.2.3 Screw Everyone! Phrases

These phrases condemn subgroups in the culture. The speaker makes use of language that reveals he or she inhabits a world that is sexist, pornographic, ethnically cleansed, racist or homophobic.

5.2.3.1 Sexual

These phrases employ a wide spectrum of sexual and anatomical references related to regular heterosexual sex. Regular dictionaries contain words like “masturbate.” The H-Dictionary involves more XXX words not fit for family audiences.

5.2.3.2 Squalid

These phrases reference physical filth or sexual fetishes. As one might expect, the majority of these phrases are in H-dictionary. The regular dictionary contains more sanitized phrases depicting squalor.

5.2.3.3 Slurs

These phrases cover cultural or ethnic name-calling. The regular dictionaries reference cultural labels of derision (“molester,””draft dodger”). The H-Dictionary references religious slurs against Jews, Christians, Italians, Irish, etc.

5.2.3.4 Homophobia

These phrases record anti-gay/lesbian sentiments. The regular dictionaries cover generics like “sissy man” and “butch.” The H-dictionaries capture more extensive anti-gay vocabulary.

5.2.3.5 Racist

These phrases target individuals of African, Asian, etc. decent. All are in the H-Dictionaries.

6 Maintaining a Flames Dictionary

The main reason for organizing flaming patterns into a hierarchical scheme of categories and phrases is to make the dictionaries easy to comprehend and maintain. Maintenance requires keeping logs on three types of phenomena.

6.1 Entry Omissions

A flames dictionary will never be complete. It is important to keep logs of patterns that the current parser misses and add them to the phrase dictionaries. Maintainers should be careful about making additions, however. Additions risk the multiplication of false positives. As mentioned above, much personal flaming revolves around language that is not conventionally associated with the language of flaming. These patterns can't be added to the flames dictionary without losing much of the overall generality of the entries. For any one context where "When will the report be out?" (See page 3 above) is a hurtful flame, there are dozens of other contexts when it has no such flaming effect (or intent). Besides capturing some linguistic generality, maintainers should also make sure that the flame patterns recorded have as much phrasal intactness as possible. The words "screw" or "screwed" have less intactness than "screw you" or "you

screwed me" as flames. While the less intact phrases may still be part of evidence for a flame, they need to be set as a weaker type of evidence in need of surrounding linguistic input.

6.2 Entry Misclassifications

After more experience with the flame parser, the maintainers may decide that individual entries are not in the appropriate phrasal categories. It is essential that there be a "living document" connected with every phrasal type that can be updated as necessary.

6.3 Category Omissions

After more experience with the flame parser, maintainers may decide that more phrasal categories need to be added. This is inevitable if the flame meter is to become more customized to individual contexts and domains of users. The categories I have developed fill a seed dictionary, with entries that are as publicly generic as I could make them. Even at this generic level, there may be glaring omissions that need to be filled in. This will become apparent once the flame parser comes to enjoy a regular set of users.

7.0 Why Monitor Flames, or Language at All?

I conclude this paper with a question that has remained in the background but not brought to the surface. Why monitor flames, or language at all? Monitoring grammar mistakes (e.g., subject-verb agreement and split infinitives) seems helpful to the writer

and considerate to the reader on the other end. Monitoring functional, purposeful, uses of language may seem to cross the line into the censorship of content.

There are two points worth making in response. First, the use of a flame meter needs to be a voluntary act. People should have their language watched only because they want it watched and feel they can benefit from the scrutiny. Second, it is simplistic to think that linguistic acts of directed aggression or raunchiness (or linguistic acts of anything, for that matter) are something a writer controls in full. Texts are among the world's most exquisite and intricate information spaces; and among the hardest information spaces to "look into" with perceptual accuracy.

As any writing teacher or editor comes to learn, even highly experienced writers can build images of their text that bear strikingly little resemblance to the actual text they have produced. Errors and lapses of various kinds are legion among writers who know better. The detection of lapses in an information space that is notorious for concealing them is the basis of grammar and spell checkers – not so much to learn spelling and grammar but to make sure writers are not blinded from using less than they know. The same applies to a flame monitoring system.

Most writers know better than to violate rules of common courtesy and social decorum in their email messages to others. But amid the large and subjectively infused range of choices that define interactive email, writers commonly lapse. They often produce language that they come to regret, disavow or ascribe to a temporary lack of control. Even if they can recover to edit their thoughts, they often find it difficult to know if their editing has been consistent or complete. Stranded islands of anger and offensiveness, ghosts from a previous draft, can lurk in shadows of the text that the eye

never catches. A system to watch flames can make sure that these ghosts are eradicated. But only if the writer wants them to be.

###