1

Multimodal Construction Grammar

FRANCIS STEEN AND MARK TURNER

1 Introduction: Massive Multimodal Data

When people meet, they invariably communicate in multiple modalities: the eyes, gestures, and tones of voice merge with the perceived affordances of the surroundings into an integrated and partially shared experience. Multimodal communication predates and contextualizes language, and extends into a series of social, artistic, and technological innovations, from dance to cave paintings, from theater to cinema, from town criers to television news. Over the past hundred thousand years, our species has grown from a few roving bands to some seven billion individuals, linked by densely tangled electronic networks. That adds up to a lot of multimodal communication, a rich treasure trove of data comparable in complexity to the information available to astrophysicists, zoologists, and geneticists. Yet the datasets scholars rely on for deciphering the hidden orders of human communication remain overwhelmingly textual. The relatively few and small corpora we have of multimodal communication often come from specialized circumstances, such as interviews conducted by experimenters in whiteroom lab settings. In this article, we explore ways to broaden the foundations of our study of human communication.

Methodologically, we need to check our hypotheses against data. In the face of small, biased, and narrow archives of data, language scientists have often relied on personal introspection to choose between hypotheses. Yet, it is well-recognized by now in cognitive linguistics that although personal intuition can be a source of hypotheses, it has weaknesses as a test of hypotheses. For example, as Dabrowska (2010) shows, judgments made by

Language and the Creative Mind.
Borkent, Michael, Barbara Dancygier, and Jennifer Hinnell (eds.).

linguists diverge from those made by the general population. She concludes that "syntacticians should not rely on their own intuitions when testing their theories."

For all these reasons—the sparseness and the biases of the data, the relative lack of systematicity, the relative lack of multimodality, the unreliability of introspective data—there has been a push in linguistics to develop corpora and methods for investigating cognitive linguistic questions through big datasets, some of them multimodal.

We hope to assist in this project for a new scientific future by deploying the resources of The Distributed Little Red Hen Lab. Red Hen (for short) is a global enterprise whose goal is to create a massive systematic corpus of ecologically valid multimodal data, along with new tools and practices to analyze this data. We record audiovisual news broadcasts systematically, and supplement the resulting dataset with other audiovisual records. This is made possible by section 108 of the U.S. Copyright Act, which authorizes libraries and archives to record and loan audiovisual news programs. We sketch here a path toward computer-assisted, statistically-assisted analyses of massive multimodal communicative data and provide some illustrations of Red Hen's current capabilities. The supporting website for this article, which archives audiovisual materials, is on the system's citation server, at **http://vrnewsscape.ucla.edu/mind/SteenTurnerMMCG.html**.

2 The Data

Red Hen's core dataset consists of the NewsScape Library of International Television News, housed and maintained securely at the library of the University of California, Los Angeles. Other datasets are in development. NewsScape consists of roughly 200,000 hours of broadcast network news, in a variety of languages. These data reveal the quick cultural creativity and extraordinary cultural variation of network news. They include roughly a billion words of timestamped closed-captioned texts, and roughly a billion words of transcripts. Red Hen ingests another hundred hours or so of global network news daily. We have begun to capture several Scandinavian channels and two channels from Spain, and we are currently working on expansion to other languages.

Red Hen has developed code for putting transcripts, where they exist, into time-stamped registration with the closed-caption text and aligning both with the audiovisual stream. Red Hen is also extracting on-screen text with optical character recognition and exploring ways to deploy speech-to-text transcription for broadcasts that lack closed-captions. Funded by a four-year National Science Foundation grant, Red Hen employs graduate students in Statistics, Computer Science, Information Studies, Communica-

tion Studies, and Political Science at UCLA and the University of Illinois Urbana-Champaign (UIUC) to develop techniques for automatic story segmentation, topic hierarchies, named entity recognition, speaker identification, and geospatial tagging. Computer vision students work on person detection, face recognition, and the characterization of visual features in facial expressions, clothing, and pose. Landscapes, buildings, vehicles, objects, and symbols are also being tagged, to permit richer description of the affordances of the environment and their systematic use in visual communication. Scholars of audio analysis examine the collection for strong emotions, vocal characteristics, and patterns of music use. Information scholars are developing new and interactive forms of data visualization. Broadcast news presents a series of hard problems to multiple disciplines, calling for transdisciplinary collaboration.

Red Hen data are highly multimodal, including speech, on-screen text, gesture (broadly defined), bodily stance, music, sound effects, graphics, and a range of other audiovisual expressions. Red Hen data are legendarily ecologically valid: human beings find TV news immersive throughout their lifespan, in all developed cultures and in all of the most-commonly studied languages. The data are broadcast daily to billions of people worldwide—for an impression, see the supporting website, which presents pictures of people in various environments attending to the news.

While the experience of watching the news is now ubiquitous, as a mode of human communication it remains a cultural novelty. The technologies and practices of multimodal news communication have undergone rapid developments over the past century—a blink of an eye compared to the history of language. They represent significant cognitive innovations: at no previous point in human history have people had the ability to assemble representations of events that to a high degree reproduce the human experience of vision and hearing. The technologically mediated communicative practices of television news present an opportunity and a challenge to our understanding of the human communicative system, inviting us to extend our study of linguistic phenomena into new multimodal forms of expression.

What do we think is happening when we watch the news? How do we make sense of the scene? Human beings have a rich experiential understanding of scenes of face-to-face "communicative joint attention." Joint attention is a human-scale scene in which some people are attending to something and know they are all attending to something and know also that they are engaged with each other in attending to it (Tomasello et al. 1995, Tobin 2008). In "communicative joint attention," people are not only jointly attending but also communicating with each other about the focus of their attention, even if the communication is very sparse, consisting perhaps of

only pointing. We use the term "classic joint attention" to refer to perhaps the most fundamental scene of communicative joint attention, in which two or a few people in face-to-face presence are not only attending to something that is directly perceptible but are moreover communicating about it in a sustained way. (Thomas & Turner 2011, Part 3.) The multimodal patterns of managing classic joint attention have been closely studied in the tradition of Erving Goffman (1967), notably in e.g. (Kendon 1990).

Watching the news is not a scene of classic joint attention, but it tacitly builds upon that understanding. The many mental spaces needed to make sense of a scene of watching network news include classic joint attention, the broadcast viewer, everyone in the viewer's environment involved in jointly watching the broadcast, everyone outside the viewer's environment involved in watching the broadcast, the staff involved in crafting the communication, the crews that handle the technology, the technology itself, the items that are the focus of the news and to which our attention is directed, and so on and on and on. This diffuse mental network would be intractable to the viewer except that we can blend its many connected mental spaces into an anchoring scene of *blended joint attention*, much of whose structure is provided by the all-important input space of classic joint attention. In the blend, we are in a scene of human-scale classic joint attention. Some of the management and communication techniques available in classic joint attention project down easily to the blend, some do not, and the blend develops emergent techniques of its own, which it is our purpose to study.

We are not deluded by this blend; contrary to the claims of Reeves & Nass (1996), it is not simply the case that we equate the media with reality. The blend integrates conceptual structure from multiple sources into a seamless and mentally manageable whole. We know, for example, that, in the full mental network, the anchor who addresses us as "you" or who says "I will show you that in a minute" and who is looking at "us" does not actually know us or see us. In the vast mental network surrounding and anchored by the human-scale scene of blended joint attention, we know that there are hundreds of agents, and people we do not see, and technological manipulation, and so on. Our attention, however, is typically focused on the meanings made available through high levels of compression achieved inside the blend. Inside the blend, we are in a congenial scene structured by the concept of classic joint attention. Accordingly, many of the linguistic constructions and words and gestures developed for classic joint attention can be projected to the blend and used directly of the blend. We do not need to invent entirely new language or gestures in order to run and understand this scene of news broadcasts. To be sure, as we will see, new or extended communicative constructions can also emerge in this blend, but they are

based on constructions we already know for our normal scenes of classic joint attention.

One of the basic research missions of Red Hen is to analyze such multimodal constructions.

3 What is a Construction?

Construction grammarians use the term "construction" in a range of ways, but in all of them, a construction is a mental packet, consisting of a form-meaning pairing that speakers of a language know. Goldberg writes,

> A construction is defined to be a pairing of form with meaning/use such that some aspect of the form or some aspect of the meaning/use is not strictly predictable from the component parts or from other constructions already established to exist in the language. On this view, phrasal patterns, including the constructions of traditional grammarians, such as relative clauses, questions, locative inversion, etc. are given theoretical status. Morphemes are also constructions, according to the definition, since their form is not predictable from their meaning or use. Given this, it follows that the lexicon is not neatly delimited from the rest of grammar, although phrasal constructions differ from lexical items in their internal complexity. Both phrasal patterns and lexical items are stored in an extended 'constructicon.' (Goldberg 1996, page 68.)

A construction might include various related elements for which traditional linguistics has various names—phonology, internal syntax, external syntax, semantics, pragmatics, and so on. Constructions can be more or less "lexically filled." For example, the "Way" construction (Israel 1996) and the "What's X doing Y?" construction (Kay & Fillmore 1999) require that the lexical item "way" and the lexical items "what" and "doing" fill the right spots in instances of the constructions. Examples of the "Way" construction include "she found her way to the market" and "he lied his way to the top." Examples of the "What's X doing Y?" construction include "What are you doing talking to me?" and "What's this coffee cup doing on my coffee table?"

The basic vocabulary we need to talk about research into multimodal constructions comes from the notion that to know a language is to know a *relational network of constructions and how they can be blended to create forms* that when expressed count for (many) other speakers of the language as belonging to the language. They count as expressions in the language because those other speakers are able to find in their own relational network of constructions some established constructions that blend to produce just those forms. By "constructicon," we will mean this knowledge of constructions and how to blend them. Broadly, knowing a language is knowing such a relational network of constructions and how its constructions can be

blended to produce forms that can be expressed. Accordingly, the two-year-old and the eighty-year-old from different geographical areas can both be thought to "know" a single language even though what they know might be considerably different. We modify our knowledge of constructions and the way they blend as we develop.

Crucially, the constructicon is multimodal. We can sum up our project as the investigation through the powers of Red Hen of multimodal constructions and how they blend. In exploring the ways in which the idea of a construction in linguistics extends and generalizes to technologically mediated forms of multimodal communication, as they have been extended to interpersonal forms of multimodal communication, we remain open to the possibility of radical discontinuities and unprecedented innovations. Our point of departure is that network news relies heavily on a carefully orchestrated system of generating blended joint attention, creatively building on a series of constructions we recognize from the analysis of linguistic communication.

4 Linguistic Constructions

One way to use Red Hen is to study her texts. These include closed-captions, transcripts, and on-screen text. In this way, Red Hen offers a dataset for garden-variety linguistic inquiry. Red Hen is for this purpose similar to other corpora, such as the Corpus of Contemporary American, the British National Corpus, and the Russian National Corpus.

As an example of how Red Hen can help us study constructions, consider Turner's (1987) analysis in *Death is the Mother of Beauty* of conceptual connections and their integration. Its data consisted of expresses using kinship terms, such as "mother." The data were acquired by tracking down remembered examples, recording adventitious encounters with expressions in speech and writing, scouring concordances and historical dictionaries and lexicons, and otherwise making use of all of the back-of-the-envelope tricks of data acquisition that characterize much of traditional linguistics. The data were recorded on 3-inch-by-5-inch index cards and stored in index boxes until they could be typed into a flat file in an account on a unix mainframe at the Lawrence Berkeley Laboratory. We now live in a different universe. A search of Red Hen finds thousands of such examples in seconds. The search produces ranges of non-literal uses of kinship terms conforming to the analyses in *Death is the Mother of Beauty*, such as "Ayn Rand is the mother of objectivism," "Virginia is the mother of presidents," and, in a usage familiar from Saddam Hussein's famous phrase, "the mother of all battles" (reportedly understood quite differently in Iraq and the United States), usages such as "Virginia is the mother of all battleground states"

(where, contextually, the battles are about US national elections) and “She is the mother of all whistlers.” One finds memorable examples such as “Pink is the mother of all colors.”

These examples can be studied or at least perused by the researcher one after another, in human time and at a human pace. But Red Hen is additionally ready to export all her hits to a comma-separated-value file, which can then be fed to a software package like R (see www.r-project.org), and subjected to statistical analysis.

The clausal construction involved in “Death is the mother of beauty” is the “X is the Y of Z” construction. This xyz construction has routine everyday uses, as in “Paul is the father of Sally.” It has been analyzed by (Turner 1991, 1998, Fauconnier & Turner 2002). xyz contains the “y-of” construction. A “Y of” expression prompts us to perform the following operations:

1. Call up an input space for the relational frame containing y (the element named by Y).
2. Construct a blended space.
3. Project from the element y selectively to create an element y' in the blend.
4. Provide for a w in the input space that will bear an appropriate relationship to y.
5. Project from that element w selectively to create an element w' in the blend.
6. Project the y-w relationship selectively onto y'-w' in the blended space.
7. Provide open-ended connectors from y' and w' in the blend. We expect these connectors to make connections at some point.
8. Expect the open-ended connector from w' in the blend to connect to something picked out by the noun phrase that will follow "of."

In an xyz construction, we additionally call up a relational frame containing x and z (the elements named by X and Z), and attach the open-ended connector from y’ to x and from w’ to z. It is possible to compose y-of constructions. That is, what follows the "of" in the first Y expression can be another Y expression, for as long as we like: "The doctor of the sister of the boss of Hieronymous Bosch."

Red Hen can of course be asked to search for examples of such constructions. One could, for example, check various hypotheses about frequency of patterns. In the Red Hen data, we find two standard patterns for y, as follows: (1) y belongs to a standard frame commonly applied to the x-z scene; y is a role connecting at least two things in that frame; x is the value of one of those roles and z is the value of the other. Examples are *archbishop* and *aunt*. (2) y is the anchor of an entrenched generic blending template used to blend together two conflicting frames. Examples are *ancestor, an-*

chor, architect, author, backbone, bane, birthplace, blood, blueprint, bottleneck, capital, cradle, . . . as in phrases like “He is the architect of our business plan and that business plan is the backbone of our operation.” But, impressively, one also finds in Red Hen very many data calling for blends of strongly conflicting frames where there is no already-entrenched y-based blending template. Examples are “The head of the C.D.C. told Congress MRSA is the cockroach of bacteria,” “Bakersfield is the Alaska of California,” and “These flame retardants are the asbestos of our time.”

Red Hen provides data for considering related constructions, of the sorts originally studied in (Turner, 1991), such as the $xy_{adjective}$ z form. When the y in an xyz conceptual pattern is a commonplace transformation of one thing into another, its form may be $xy_{adjective}z$, so "Language is the fossil of poetry" may be expressed as "Language is fossil poetry." When the y-w conceptual relation is a part-whole frame relation, the form may be $xz_{adjective}y$, so "Las Vegas is the Monte Carlo of America" may be expressed as "Las Vegas is the American Monte Carlo." Red Hen offers up great ranges of such constructions related to the xyz construction, such as “Westerns are back in the TV saddle,” “She is gymnastics royalty,” “He is passing the ethical buck,” and “Iran could be this year’s Japan.”

Investigations along these lines are in principle no different than investigations run on any text corpus, although for some purposes Red Hen may offer some advantages derived from content or ease of search. Red Hen to this extent takes her place alongside text corpora.

But because Red Hen is multimodal, we can instantly look for *visual* correlates of the xyz construction. Network news, to give viewers an idea of what some event is about, frequently position a photograph or footage of the new event next to a photograph or footage from some old, well-known event, and this presentation is not to be understood as indicating that the two represented scenes combine or connect in reality. For example, when Anders Behring Breivik exploded a car bomb in front of government offices in Oslo in 2011, several news channels showed his face alongside that of Timothy McVeigh, the 1995 Oklahoma City bomber, visually conveying the message that Breivik is the Timothy McVeigh of Norway.

5 “Errors”

As Hofstadter and Moser (1989, 185) observed, “the study of speech errors and action slips can reveal a great deal about the hidden organization of the minds that produce them.” Because so much of the speech in news broadcasts is improvised, extemporaneous, and live, as opposed to written or edited or both, Red Hen’s data may have advantages in containing linguistic “errors” that might serve as windows on grammatical processes. For exam-

ple, it is a natural hypothesis that when there are rival forms available for prompting for a meaning that the speaker wants to express, and moreover those forms are related in a way that makes it easy to produce a blend of both, "errors" might result, as in "This is my future wife-to-be" (Cohen 1987). But if we are right in our hypotheses about the production of such errors, then we should be able to predict their relative occurrence, probabilistically. We might have to listen for an impractical amount of time in order to acquire enough data to test our hypotheses, and of course these errors are likely to be edited out of written data or even unconsciously suppressed during human transcription. But Red Hen has the actual performances and increasingly has machine transcription and speech-to-text transcription that may preserve the errors; in addition, the audio recordings are instantly available for systematic verification.

Let's ask Red Hen to do it, as follows: There are many patterns related to the xyz construction. Let us return to mathematics and create an example of an xyz clause: "This number theory problem is the Mt. Everest of mathematics." The w, presumably something like *challenging mountains to be climbed,* is not mentioned. We have the following related patterns:

1 xyz itself, "This number theory problem is the Mt. Everest of mathematics." w is not mentioned.
2 x is y, "This number theory problem is Mt. Everest." Neither w nor z needs to be mentioned.
3 x is the z-equivalent of y, "This number theory problem is the mathematics equivalent of Mt. Everest." (This is a case, for us, on the cusp: We could say equally easily, "This number theory problem is the mathematical equivalent of Mt. Everest.")
4 x is the equivalent of y, "This number theory problem is the equivalent of Mt. Everest."
5 x is the $z_{adjective}$-y, "This number theory problem is the mathematical Mt. Everest."

But 4 and 5 are very close. Since $z_{adjective}$-y is a noun whose head is y, it is easy to slide it into the noun-phrase position for y in 3. Do we get "errors" where the speaker produces "x is the equivalent of $z_{adjective}$-y," like "This number theory problem is the equivalent of the mathematical Mt. Everest"? We asked Red Hen, and found just such an "error," produced by a high-end car dealer who was engaged in an interview with a TV reporter as they walked amongst some spectacular antique cars that were to be auctioned at a legendary event the next day: "That car is the equivalent of the automotive Mona Lisa." (See the supporting website.) Any editor would have corrected that sentence before publication. The accepted alternatives are: "That car is the Mona Lisa of automobiles" (pattern 1), "That car is the Mona Lisa"

(pattern 2), "That car is the automobile/automotive equivalent of the Mona Lisa" (pattern 3), "That car is the equivalent of the Mona Lisa" (pattern 4), and "That car is the automotive Mona Lisa" (pattern 5). And yet, what would have been regarded as an error in written language may be much more acceptable in multimodal communication: the overdetermining "equivalent" in this sentence provides an additional cue to the hearer to construct a blend, and the extra cue may be communicatively useful. Red Hen can accordingly be used to find and analyze not only familiar linguistic constructions but also context that might lead to the production of what are routinely regarded as errors, and, given the massive amount of data in Red Hen, to test predictions of statistical patterns of errors.

6 Co-Speech Gesture

The study of co-speech gesture has flourished in the last few decades, and Red Hen offers a great wealth of data because all of the text is tagged with thumbnails that one can click to see a recording the actual performance of the utterance, including gesture. It is easy to think of indefinitely many investigations of co-speech gesture for Red Hen to run. For example, although the temporal adverb "previously" is etymologically connected to movement along a path, it is probably not analyzable as such to most current speakers of English. How do they gesture when they say it? We asked Red Hen to find in a very limited date range examples of "previously" at the end of a sentence. Red Hen found many, several of which have gestures, even when anchors, trained out of hand gesturing, use the word. Senator Kelly Ayotte on CBS Face the Nation for 2013-02-24, says "You had the Secretary of Education on previously," and nods her head laterally to her left when she says the word. (See the supporting website.) Red Hen hopes to help develop machine recognition of gesture and annotation of text for co-speech gestures. This machine recognition is built on training from clips that have been hand-tagged by human experts. Naturally, Red Hen data might be better or worse for specific purposes in gesture study. Those who prefer elicited data from controlled experimental conditions might have methodological concerns over using observational data, but, on the other hand, Red Hen data has the advantage of being completely free of the corruptions that come from demand characteristics of experimental data involving human beings or the strangeness of elicited conversation under lab scrutiny. Those who prefer to study spontaneous, untrained gesture may have qualms about studying the gestures of such trained performers as news anchors and reporters, but, on the other hand, first, a great deal of Red Hen data consist of spontaneous, untrained performance by people who happened to be recorded in public places, many of whom were presumably unaware that they

were being recorded, and, second, perhaps there is much to be learned about communication by studying the performances of highly expert communicators, who, for all their training, often land in the position of extemporizing without a script or without even warning of the breaking news story.

We offer as an example of co-speech gesture that can be studied within Red Hen the performance of US Secretary of State Hillary Clinton while testifying before the Senate Foreign Relations Committee, chaired by Senator John Kerry, on Capitol Hill in Washington, DC, March 2, 2011. (See the supporting website.) Her testimony was recorded and broadcast by Russia Today, whose main news broadcast is systematically ingested by Red Hen. Of course, politicians use a vocabulary of prepared gestures (e.g. closed rather than open-finger) that is quite distinctive relative to everyday gestures, but then again, the study of prepared performance in gesture is interesting in its own right, and Red Hen has massive gestural data from non-politicians or, more generally, non-performers. One research project using Red Hen would be to contrast the practiced performers with others. In a way, everyone is a practiced performer; the question is instead what they have practiced for.

The RT anchor frames the situation and introduces the topic to the viewer:

> Hillary Clinton says the US is losing an information war to foreign media outlets, including RT. This week the Secretary of State asked Congress for more cash to step up America's efforts to get its message across.

We begin by recognizing viewpoint: the situation defines Clinton in the role of a Secretary of State arguing for enhanced funding of her Department. Her public performance thus embodies an enduring institutional logic, one that is reflected in and reinforced by the details of the visual setting: the microphones, the panel of senators, the sheer presence of the media. This acknowledged role play does not vitiate the force of her arguments; rather, her performance has the force of the entire State Department, underlining how seriously the institution views the development of competing global television news outlets.

Against the backdrop of a silent and grim Clinton, the on-screen text flashes,

> CLINTON: US LOSING GLOBAL INFO
> WAR AS MULTI-POLAR REALITY BITES

The multimodal blend assembles these two sources of information—the facial expression and the on-screen text—into a single, integrated interpretation: Secretary Clinton is the embodiment of the State Department and the entire US, and she expresses, in a compressed form, the emotional meaning of a loss, coupled with an aggressive determination to attempt to face the

facts and avoid further losses. The unfocused gaze and drawn mouth suggest sadness and resignation, while the clenched jaw hints of a determination to fight back. The human visual system interprets faces with extreme speed, and without needing to translate the features into verbally expressed concepts; only a slowed-down analysis allows us to identify the elements of the crossmodal blend.

Yet Clinton's clenched performance is effortlessly contained by the triumphal ethical superiority of the RT reporter, who melodramatically proclaims,

> War declared. The US is now officially in an information battle with foreign media, which provide alternative views on world news -- views which often run in contrast to the coverage of events by the US mainstream media.

Before the Secretary's testimony begins, RT has already reconfigured the loss of US dominance as progress, a positive development facilitating multiperspective understanding.

Gesturally, Clinton opens by metaphorically taking hold of the entire concept in a precision grip, staring intently with wide-open eyes on her unseen interlocutor, declaring "We are in an information war." Casting her glance briefly aside, indicating thought, she angles her wrist and hand to point towards herself, saying "And we are losing that war; I'll be very blunt in my assessment," as she tilts her head slightly to the right, appearing less imposing or threatening, and bobs her head up and down with a slight smile, as if eagerly eliciting agreement.

In the space she creates with these simple gestures, Clinton situates herself as the US in the center of a symbolic space. With an open hand she swings her right arm in an upward sweeping gesture, opening up this space to the competition in the global war of information; her gaze pointedly fixed straight ahead, she extends her arm fully to the side and down in an expansive movement, freezing her shoulder in the down position, saying, "Al Jazeera is winning." The gesture masterfully conveys at once an unspoken acknowledgment of the strength, significance, and enduring presence of the opponent and his unwelcome status; he is not granted the inclusive grace of even a glace, a brief moment of symbolic eye contact. After a brief, pregnant pause she raises her extended hand slightly, opening up the space further as she expands the list: "The Chinese have opened up a global English-language and multi-language television network," nodding at each word to convey the brute fact of the audacity of the encroachment. Finally, "The Russians have opened up an English-language network. I've seen it in a few countries, and it's quite instructive." Closing her eyes for a moment in the cognitive equivalent of a black screen transition, she raises her arm and indicates a retreat by gesturing repeatedly towards herself: "We're cutting

back, the BBC is cutting back." The Secretary herself is gesturally mapped as personifying the entire Anglo-American alliance, graphically under threat by other perspectives. This bodily language, as Seana Coulson has shown (this volume), facilitates the rapid construction of the complex cognitive models that Clinton is compressing into a few minutes of testimony.

As viewers of RT, we watch all this nested within an outer frame: in an ironic twist, the Secretary is testifying that the US is losing an information war precisely to the channel in which we are watching her testimony. Watching RT, we become allies of alternative sources of information, positioning us on the disruptive but winning side in the information war.

7 Audiovisual Correlates of Linguistic Constructions

Cognitive linguists routinely study basic mental operations and phenomena that are not exclusive to language but that are deployed in language and leave their mark on its structure: mental space phenomena, conceptual integration, categorization, image-schematic structuring and transformation, fictive motion, force dynamics, viewpoint phenomena, scanning . . . Since the news deploys other modalities than speech and text, it is an obvious project to look for the ways in which these basic mental operations and phenomena are deployed in those other modalities.

Many words in many languages prompt us to build conceptual integration networks containing relations of counterfactuality, such as "safe," "accident," "dent," "mistake," "Good thing . . . ," "Too bad, . . .," and so on. The news presents counterfactuality so routinely that it is surprising that the stereotype of the news is that it presents "what's happening." The news has standard routines for audiovisual prompting of the counterfactual. One of the most straightforward of these standard routines is to show a visual prompt for the actual counterfactual scene. For example, a young woman who stars as a student in a based-on-a-true-story film about a high-school teacher who was fabulously successful in teaching mathematics to inner-city poor youth reports, "I thought if only I had come to this school, if only I had him for a teacher, I would have learned math," and the TV news immediately shows a picture of a young woman dressed as a student standing with the legendary teacher, so we can map the speaker onto the student. (See the supporting website.) A woman at the beauty parlor says "If only I had some shoes", and the TV news immediately shows a vast array of shoes. (See the supporting website.) It is easy for us to imagine the counterfactual, but not to *perceive* it. Television is using its standard technique in these cases of moving an operation that is normal to human imagination into actual human perception. The bizarre perceptual phenomenon does not seem bizarre to us at all because it is a familiar imaginative phenomenon.

Such data, in which the news visually presents a representation of the counterfactual scene or counterfactual elements needed for constructing the blend, are vast and often subtle and nuanced, falling into a variety of types. Consider the word "detour," which prompts us to construct one space with one path A on which something travels and another space with another path B on which the same thing travels. When we blend these so as to take both paths but travel on A, then the counterfactual relationship between travel on A and travel on B is compressed to *absence of travel on B*. Language for expressing this blend includes "direct," "quick," and so on. If we blend the same inputs so as to take both paths but travel on B, then the counterfactual relationship between travel on A and travel on B is compressed to *absence of travel on A*. Language for expressing this blend includes "detour," "delay," and so on. Red Hen offers a specific case in which police traveling to an island took path B by inflatable boat, and during the period of the travel, scores of young people were murdered on an island not far from Oslo, Norway. In this presentation, the "detour" causes a "delay," used as an invitation to the audience to construct a network with a counterfactual connection between the two journeys. Without explicitly accusing the police, the news is designed to lead its viewers to contemplate the alternative scenario and to compare the two outcomes. The delta of this operation is additional deaths: taking the "shortest route" would accordingly have "saved lives." All during the linguistic presentation, the news is showing a satellite image of the geographical area, with the diagrammatic trace of the possible and the actual travel by the police. The dynamic travel that in fact did not happen is shown first, followed by the route actually taken (see the supporting web site). The words "Dette er den korteste veien til Utøya" 'this is the shortest route to Utøya' and "politiet valgte å kjøre denne omveien" 'the police chose to take this detour' are accordingly paired with visual presentations for the blend that prompt for the activation of the different input spaces, the counterfactual relationship between them, and the compression to features like *absence, detour, delay*, and so on.

This depiction of the counterfactual exemplifies how the news uses the multimodality of the presentation to prompt for the understanding that we linguists usually analyze by inspecting the linguistic phenomena only. Red Hen would seem to be a natural dataset for such investigations. Presumably, for example, one could investigate gestural prompts for the construction of counterfactuality and blending to produce emergent structure.

Let us consider a specific grammatical construction for which there was originally an implicit hypothesis about its natural alignment with multimodal presentation. Nikiforidou (2010 & 2012) analyzes the role of blending in a construction she calls "Past tense + proximal deictic," with emphasis on the cases where the proximal deictic is "now." This well-known construc-

tion is also broadly analyzed in (Dancygier 2012, chapter 7; Sweetser, this volume). The preferred patterns are "was/were + now," as in "It was now possible . . ." and, for a non-copula verb, "now + past tense," as in "He now saw that . . ." Nikiforidou provides "a detailed blueprint of the blending mappings cued by the [past + proximal deictic] pattern" (2012, 177). Essentially, the pattern calls for a blend of viewpoints, in which our overall understanding is stage-managed from the point of view of a narrator but some self or consciousness located in a previous time is contextually available and prominent, and the events experienced in that previous time are to be construed "from the point of view of that consciousness, as that character's thoughts, speech or perceptions" (2010, 266). The blended viewpoint takes on elements of different perspectives and compresses a time relation. The mental space of the narrator's condition is still the mental space from which the narrated space is accessed and built up, but the experiential perspective comes from inside the narrated events. There is considerable emergent structure in the blend. In the blend, it is possible to have not only knowledge that is available only at a distance but also to have the experience, perception, and realization available only up close.

In a study of the British National Corpus, Nikiforidou shows that this is a highly productive construction, even outside of literary genres.

Presciently, Nikiforidou writes that the grammatical pattern has the "effect of zooming in on the events" (Nikiforidou 2012, 180). Let us consider "zooming in." A human being can shift focus and attention and can locomote or lean forward or back in order to change the angle that objects subtend in the visual field, but, absent assisting technology, a human being cannot zoom in on or zoom out from an actual percept—that is, a human being cannot, without changing the location of the head, change the angle that objects subtend in the visual field. Yet, it is easy to "zoom" in visual imagination, as anyone can demonstrate. In your mind's eye, picture some landmark near where you live—a familiar building, a bridge, a body of water, for example—or picture a person. Now, in visual imagination, zoom in. Now zoom out. Zooming seems to be a very common and easy manipulation in imagination. Perhaps imagination works this way because we do have the actual experience of an image in the visual field getting "larger" as we move toward it or "smaller" as we move away from it, and we can of course do time-scaling in imagination because of blending, so zooming in imagination is available as a product of blending over experience, but not projecting to the blend the movement of the viewer. Does the news adopt this mental functionality of the zooming imagination by using a photographic visual-field analogue of mental zooming? That is, does it move an operation from imagination into perception? Yes. For screens that result from camerawork, we have available a mental blend in which our eye is

fused with the camera lens or with the through-the-lens viewfinder. Now, when the camera zooms—which it can do because it has a lens array that our eyes do not have—our vision, in the blend for the news, *zooms*. So the next question is: When we are in a news scene of narration and the narrator uses the *past + now* construction, does the camera zoom in on the self or consciousness that had the experiences?

Yes. Or rather, that is what we often find in Red Hen. There is a hitch in providing this *past + now* visual zoom, because the narrator speaks *at one time* about a consciousness *at a previous time*. That is a mismatch. The consciousness and its experiences are not available in the narrator's immediate environment, or indeed in any of the mental spaces we have for considering the production and broadcast of the narration. How do we deal with the mismatch? The news production team must provide some suitable prompt for that consciousness in the past, but it isn't with them. There are several ways to resolve the mismatch. The three most common appear to be (1) have the person who is coreferential with the consciousness we are narrating re-enact the events, with the appropriate setting and staging and so on, and film that scene; (2) find archival still photos of that person at the time and present them, perhaps, e.g., with a Kens Burns effect, as the narrator uses the *past + now* construction; (3) find historical film footage containing the person and run that footage that as the narrator uses the *past + now* construction. One can, of course, do all three. For all three of these expedients, one can zoom in on the images. Of course, narrators on the news usually just narrate, without using this extra production, but such zooming production is easy to find in Red Hen. Nikiforidou did not explicitly predict these data, but implicitly she did, and the data support her intuition of what is going on mentally.

Here are two examples (see the supporting website for the clips). In the first, the narrator is telling the story of Kim, who as an adult had to deal with her mother's continuing to invest in what Kim came to understand was a fraudulent business scam. While the narrator uses phrases like "Kim now saw . . ." and "Kim now wondered . . .," we see scenes of what must be a more recent Kim balancing a checkbook, shaking her head, doing sums, comparing documents, and so on. Kim is re-enacting her consciousness and behavior from the past. The production team uses camera zoom and jump cut to get closer to the image of Kim's re-enactment of her scenes of realization as the narrator uses the *past + now* construction.

Nikiforidou writes of the linguistic construction,

> In blending terms, ... resolution of (apparent) conflict is often achieved through the mechanism of compression, whereby elements that are conceptually separate in the input spaces are construed as one in the blended space. The construction at hand, I suggest, cues a particular kind of com-

> pression, namely compression of a time relation. The dynamic, continuously updated character of such blending networks renders them particularly suitable for representing meaning in a narrative, where formal clues may often give conflicting instructions even within the same sentence (as is the case with FIS [Free Indirect Speech]). (2012, 179)

We see in the news the audiovisual analogue of this blending prompt: in the news, we have input spaces in which the narrator and the production team have a viewpoint on the past and on the orchestration and arrangement of all the mental spaces involved in the network of the narrative, and we have a particular input space that has Kim and her experiences, and we are prompted to create a blended space that has projections from both the orchestrating viewpoints and Kim's viewpoint, creating there a great compression not only of viewpoint but also of time.

One of the most interesting and common uses of *past + now* occurs when the narrator and the narrated consciousness are coreferential. We find many such examples in Red Hen, such as a documentary on the Pentagon Papers, in which Daniel Ellsberg, who leaked the Pentagon Papers, is narrating in advanced age his exploits in the period 1967-1971. (See the supporting website.) As he narrates, we see historical footage of several presidents discussing the Vietnam conflict publicly, along with historical footage of the war. But we also see, in montage, close-ups and zooms of Ellsberg at the time. Perhaps these scenes are acted, using an actor to play the young Ellsberg—the audience is not told. In montage, as the historical footage runs, we are treated to close-ups of fingers—which we are to take as belonging to Ellsberg—running over the typed words of pages of the Pentagon Papers. We see "Ellsberg" opening volumes, closing them, shifting them, inspecting the documents, all in great close-up, and using various zooming techniques. Ellsberg the narrator is using the *past + now* construction: "I now saw that [President Lyndon] Johnson was continuing a pattern of presidential lying." We construct a viewpoint blend in which the viewpoint of the orchestration of the entire narrative is projected from the mental space with Ellsberg the narrator but the viewpoint on the experience of realizing what was happening is projected from the space of Ellsberg as he reads the Pentagon Papers for the first time. There is extraordinary emergent structure in this blend, including Ellsberg's ability to speak for his young self in a way that probably would not have been available to him at the time, and of course and enduring, manufactured, compressed character for "Ellsberg" the man: young Ellsberg and old Ellsberg are of course extremely different things, but the analogies between them, including analogies of viewpoint, can be compressed to a characterological unity in the blend.

Just in passing, we acknowledge that it is possible to deploy not only *reinforcing* or *supplementary* linguistic and audiovisual constructions but also

conflicting constructions. This is a standard technique of humor and various forms of entertainment. It is also a technique of news broadcasts whose brand is based in an explicit and reliable political or ideological stamp. These shows routinely ridicule the opposition by providing footage of the opposition with accompanying audiovisual constructions designed to make them look stupid.

8 Novel Broadcast Constructions

Since new constructions arise by blending existing constructions and conceptual arrays, there is really never any construction that is truly "novel." Human cognition and expression is highly conservative, although what we focus on is often the emergent structure. Accordingly, when we discuss "novel broadcast constructions," we do not mean to indicate that they are not based on antecedent constructions. On the contrary, everything new human beings contrive is deeply based in established input spaces. Otherwise, it would be not novel but rather unintelligible.

The power of broadcast news to use so many modalities simultaneously, not as a combination but rather as a system that different disciplines have tried to approach as a linear sum, produces a wealth of partly-novel broadcast constructions. Consider, for example, the extraordinary work of news broadcasts to talk about the future, as in weather forecasts. What is required in this case is amazing compressions over not only time, space, causation, and agency, but also possibility. Some of the audiovisual prompts for these compressions are spectacular but look utterly natural. For example, in weather reports about impending meteorological reports such as hurricanes, one can find the usual linguistic expressions, such as "likely," "perhaps," "threat," and so on, both in the speech of the weather forecaster and in the on-screen text. Graphics often accompany these verbal reports, and the graphics can be extremely sophisticated while striking viewers as entirely ordinary. It is common, for example, to display, using compression of time and space, the future of a hurricane. One sees representations on a map that we are to take as prompting for a conception of the path of the hurricane, and also the hurricane's size, but there is more: As the path moves into the future, there is an expanding width to the cone of incidence for the hurricane, not because the radius of the hurricane itself is expanding, but because our ignorance of the probable location of the hurricane is increasing. The conical graph on the map compresses not only time and space but also epistemic stance. We seem to have here a visual correlate of an evidential marker. (See website for an example.)

9 Conclusion

The multimodal dimensions of human communication present a rich field of discovery and insight that extends existing linguistic theories into other modalities, providing new opportunities for systematically testing and validating intuitions. The different modalities of a communicative act are not simply redundant, but provide new aspects of meaning that a multimodal analysis can uncover. Nor is the construction of meaning across modalities mechanically additive; rather, meanings emerge as crossmodal blends that rapidly synthesize selected features of the information into new wholes. To understand human communication, we must develop a new facility for understanding the grammar of multimodal meaning construction.

The Red Hen project aims to integrate a multimodal data collection of global television news with data enhancement, multimodal data mining, and cognitive analysis of multimodal communicative practices. Television news represents a uniquely available dataset of common communicative practices, with the additional attraction that it has been deliberately crafted by professional teams to grab attention, convey complex meanings in a rapid and compressed format, and persuade by implication. The way in which television succeeds in tapping into ancient interpretive structures in our cognitive system makes its detailed examination particularly revealing.

References

Cohen, Gerald. 1987. *Syntactic Blends in English Parole*. Frankfurt, Bern, New York: Peter Lang.

Coulson, Seana. 2013. "Cognitive neuroscience of creative language." This volume.

Dabrowska, Ewa, 2010. "Naive v. expert intuitions: an empirical study of acceptability judgments," *The Linguistic Review* 27 (2010), DOI 10.1515/tlir.2010.001)

Dancygier, Barbara. 2012. *The Language of Stories: A Cognitive Approach*. Cambridge: Cambridge University Press.

Dancygier, Barbara & Eve Sweetser, editors. 2012. *Viewpoint in Language: A Multimodal Perspective*. Cambridge UK: Cambridge University Press.

Fauconnier, Gilles & Mark Turner. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.

Fillmore**,** C. J. 2006. Articulating Lexicon and Constructicon. Presentation at the Fourth International Conference on Construction Grammar, Tokyo, Japan.

Fillmore, Charles J., Russell R. Lee-Goldman, Russell Rhodes. 2011. "The FrameNet Constructicon." In Boas, H.C. and Sag, I.A. (eds.) *Sign-based Construction Grammar*. Stanford: CSLI Publications.

Goffman, Erving. 1967. *Interaction Ritual: Essays in Face-to-Face Behavior*. Chicago: Aldine

Goldberg, Adele. 1996. “Construction Grammar.” In Brown, E.K & J. E. Miller, editors, *Concise Encyclopedia of Syntactic Theories*. New York: Pergamon.

Hofstadter, Douglas & David J. Moser. 1989. “To Err is Human; To Study Error-making is Cognitive Science." *Michigan Quarterly Review*, Volume 28, number 2, pages 185--215.

Israel, Michael. 1996. “The *Way* Constructions Grow.” In Goldberg, Adele, editor, *Conceptual Structure, Discourse and Language*. Stanford: CSLI. Pages 217-230.

Kay, Paul & Charles Fillmore. 1999. “Grammatical Constructions and Linguistic Generalizations: The What's X Doing Y? Construction.” *Language*, Vol. 75, No. 1., pages 1-33.

Kendon, Adam. 1990. *Conducting Interaction: Patterns of behavior in focused interactions*. Cambridge: Cambridge University Press.

Nikiforidou, Kiki. 2012. “The constructional underpinnings of viewpoint blends: The *Past + now* in language and literature**.”** In: B. Dancygier & E. Sweetser (eds.), *Viewpoint and Perspective in Language and Gesture*. Cambridge: Cambridge University Press. pages 177-197

Nikiforidou, Kiki. 2010. “Viewpoint and construction grammar: The case of *past* + now.” *Language and Literature* 19(2) 265-284.

Tobin, Vera. 2008. “Literary Joint Attention: Social Cognition and the Puzzles of Modernism.” (unpublished dissertation.)

Tomasello, M. 1995. “Joint Attention as Social Cognition.” In Moore, C. and Dunham, P. (Eds.) *Joint Attention: Its Origins and Role in Development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Turner, Mark. 1991. *Reading Minds: The Study of English in the Age of Cognitive Science*. Princeton, New Jersey: Princeton University Press.

Turner, Mark. 1998. “Figure.” In Cristina Cacciari, Ray Gibbs, Jr., Albert Katz, and Mark Turner, *Figurative Language and Thought*. New York: Oxford University Press, 1998.

Turner, Mark. 1987. *Death is the Mother of Beauty: Mind, Metaphor, Criticism*. Chicago: University of Chicago Press.