Improvements for A-F Versions 0.97[1]

A. D. Forbes –1 September 2012

0. Improvements Needed for A-F

We have long worked to correct shortcomings in our database. Our plan was to address the major ones before releasing version 1.0. Having reached version 0.95 in August 2010, the next release ideally would be version 1.0. But, health and other matters have prevented us from getting all the needed changes in place for version 1.0. Hence, we have interpolated a version 0.97. Here is a list of the four problems that we have addressed for version 0.97:

1. ***Text Type Rationalization***—Regrettably, the text types were put in place three decades ago without consultation and proper design. As a result, many of the current text types are nothing of the sort. Further, many are subsets of others. Several amount to unanalyzed ragbags. The three flavors of "Other" (**D**, **E**, and **H**) and "Oracle" (**O**) come to mind. In section 1, I briefly describe improvements that we have made. The text types remain a concern, but the changes correct some of the major deficiencies.

2. ***Periphrastics***—Our handling of periphrastics in v. 0.95 was incomplete and the licensing relations were inconsistent. We have resolved these issues.

3. ***Polysemic Cue Phrases***—Previously, we addressed the matter of the polysemy of כִּי. But all frequent cue phrases exhibit polysemy. We have put in place the apparatus needed to code for cue phrase polysemy as we get into discourse analysis and have applied it to דִּי/אֲשֶׁר.

4. ***Verb "Semantics/Valence"***—The semantics field was originally introduced for nouns in about 1985 to assist the computer as it parsed the text. Having the position available in the feature vectors of verbs, we decided to assign values for them as well. The values assigned are very hodgepodge. For this

[1] This note includes multiple references to internal representations that will be mysterious to readers. Nonetheless, it should provide insight into the changes made to arrive at version 0.97.

release, we have corrected the most egregious assignment: labeling the passive verbs (redundantly) as having "passive" semantics/valence.

1. Text Types

1.1 Improving Our Text-Type Assignments

The field in our database that we long called "genre" was renamed "text type" a while back. We have long been uneasy regarding the quality of the assignments coming under that rubric. Progress in improving our classifications has been hampered by the unfortunate fact that no one seems to have a clear idea as to how to define genres or text types operationally. As Lee (2001)[2] puts it: "…the term *text type* … can be used in a vague way to mean almost anything." Or consider Santini's contention[3]:

> "We can say that, at least so far, almost everything in the automatic genre identification research field is fuzzy, slippery, unstable, flexible (especially the notion of 'genre' and the terminology), and conditioned by the computational cost of extracting relevant features."

For A-F version 0.97, we have cleaned up the present system manually. The results are still unsatisfactory, but they are a considerable improvement over the original categories which were based on no underlying theory, exhibited wide differences in granularity, and were unacceptably inconsistent.

1.2 What Has Been Done with Text Types in Version 0.97

These improvements have been introduced:

1. Nathan's parable (**U**) has been reclassified as H-to-H narrative in speech (**n**) since the original label described a participant, not a text type.
2. Aside from a few text portions assigned to other text types, the vast majority of Moses' Torah (**M**) was reassigned to H-to-H instruction (**i**).
3. Speech-in-dialog (**h**), divine soliloquy (**X**), and human soliloquy (**x**) segments have been allocated to actual text types, freeing up symbols **h**, **X,** and **x**. (For repurposed uses, see below.)

---

[2] David YW Lee, "Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle," **Lang. Learning & Tech.**, **5**(3): 38, 2001.

[3] Marina Santini, "State-of-the-Art on Automatic Genre Identification," **ITRI-04-03 report** (Brighton), January 2004, p. 22.

4. Two new text types have been introduced:
    i. Curse— **C** (D-to-H)[4] and **c** (H-to-H).
    ii. Situation— **U** (D-to-D) and **u** (H-to-H).[5]
5. A new exchange pair has been introduced: *divinity-to-divinity* (D-to-D). This includes divine "self-talk," AKA, soliloquy. Four text-type labels suffice:
    i. Instruction (**V**).
    ii. Request/question (**k**).
    iii. Prediction/promise (**X**).
    iv. Situation (**U**).
6. Eight low frequency exchange participants have been segregated with no text-type labels supplied:
    i. Angel ↔ Human (**d**).
    ii. Clay → Human (**1**).
    iii. Donkey ↔ Human (**2**).
    iv. HolyOne → Human (**3**).
    v. Satan ↔ God (**4**).
    vi. Snake ↔ Human (**5**).
    vii. Spirit → God (**6**).
    viii. Tree ↔ Tree (**7**).
7. The table shows the symbol assigned to each recognized combination of exchange-pair (column) and text-type (row). To implement all this, almost 21,000 changes have been made. (New exchange pair/text type combinations have their internal representations enclosed in brackets.)

We have not yet begun installing the *Situation* text types for Divinity-to-Human ([A]) and Human-to-Divinity ([a]). We expect these to be allocated from present rag-bag text types D and E.

[4] Formerly, **C** was incorrectly termed "blessings." This was an instance of bookkeeping gone awry. The set of passages has been purified so it now contains 'only' curses.

[5] *Situations* are also known as *States of Affairs*, *SoA*s, in several structural-functional linguistic theories. See Chapter 8 "Representing Situations" of C. S. Butler, *Structure and Function, Part I. Approaches to the simplex clause*, (Amsterdam: John Benjamins, 2003). A typology of SoAs is there presented, along with the features and their values that allow recognition of various subtypes of SoAs. I conjecture that prospecting through the "*other*" text types (**D**, **E**, and **H**) will yield a good crop of SoAs. This text type has not yet been propagated through our data.

Text Type ↓	Exchange Pair				
	Author → Reader	Divinity → Divinity	Divinity → Human	Human → Divinity	Human → Human
Title	T				t
Genealogy	G				
Narrative	N				n
Quarrel				Q	q
Accusation			P		
Judgment			J		j
Lamentation				L	l
Instruction		[V]	I		i
Request		[k]	K	R	r
Supplication				S	s
Blessing				B	b
[Curse]			C		[c]
Prediction/Promise		[X]	Z	Y	y
Woe and Dirge			W		w
Prophesy					o
Greeting					f
Praise					z
Wisdom					v
[Situation, "SoA"]		[U]	[A]	[a]	[u]
Oracle			O		
Other			D	E	H

	Atypical Exchange Participants
d	Angel ↔ Human
[1]	Clay[1] → Human
[2]	Donkey[2] ↔ Human
[3]	Holy one[3] → Human
[4]	Satan[4] ↔ God
[5]	Snake[5] ↔ Human
[6]	Spirit[6] ↔ God
[7]	Tree[7] ↔ Tree

[1]- Isa 45:9b.
[2]- Num 22:28-30.
[3]- Dan 4:11-14, 20; 8:13b-14.
[4]- Job 1:7, 9, 10, 11; 2:2, 4, 5.
[5]- Gen 3:1, 4, 5.
[6]- 2 Chron 18:20, 21.
[7]- Judg 9:8-15.

1.3 What Still Needs Doing as Regards Text Types

We need to examine:

1. Are the text types mutually exclusive? Are not some subsets of others?
2. Are the text types operationally differentiable? For example, how do *request* and *supplication* differ?
3. Might the *divine accusation* TT (**P**) be moved up into the *quarrel* row?
4. Can the "*Other*" TTs be redistributed into present TT cells or into newly introduced rows, such as *situation*?

5. Shouldn't Oracle (**O**) be redistributed into other cells, freeing up its code? (Subtypes of "Oracle" may need to be introduced.)
6. Several empty cells attract suspicion. For example, is it really the case that no human ever accuses another human? And do humans never praise any divinity? Do deities really never bless humans? And so on…
7. Should we replace *divinity* with *celestial being* to include angels, holy ones, spirits, and The Satan under the D-to-D, D-to-H, or H-to-D headings?

All this will involve much work, involving a certain amount of circularity in that reasoned assignment of text type presupposes an adequate discourse analysis, but it is discourse analysis that we are working on the text types to prepare for.

2. Periphrastics

Both Hebrew and Aramaic periphrastics have been rationalized. Many more periphrastics have been identified and the 'auxiliary'-followed-by-participle licensing relation is now *modification* (in place of the former *join* relation). We have parsed 24 additional periphrastics, mostly inverted. We have also changed 170 *join* (**j** and **J**) licensing relations to *modify* (**m** and **M**).

3. Cue-Phrase Polysemy

To allow handling of the rampant polysemy of cue phrases, many of their vectors have been aligned to parallel our practice for the seven senses of כִּי. For this form, the POS superset is almost always J (= *conjunction*), its family is b, and its vector template is thus ?Jb.??+. Specifically, we have:

Senses of כִּי	Vector
because	?J b b??+
but	?J b e??+
that	?J b t??+
although	?J b a??+
when	?J b w??+
if	?J b i??+
surely	?? m ???+

For the (at present) three senses of דִּי/אֲשֶׁר,[6] we have:

Senses of דִּי/אֲשֶׁר	Vector
which	?? r ???+
because	?J r b??+
that	?J r t??+

In the future, the many senses of the cue phrases will be similarly vectored.

4. Verb "Semantics/Valence"

In version 0.95, 3,488 segments are identified in the semantics/valence field as being *passive*. Since passivity is signaled in the second position of the feature vector, this is redundant. Further, its presence hides items so marked from search by verb semantics/valence. Hence, we have replaced the passive flags ("p") by one of the other available flags, of which there are the eleven shown in the table.

code		code		code	
a	attitude	m	movement	s	stative
d	destruction	o	ditransitive	u	utterance sans אמר
j	transitive	p	passive	y	"say" [אמר]
k	intransitive	r	caused motion	z	estimative [קרא]

We proceeded as follows:

1. Form a special version of the dictionary wherein all verbs with passive "semantics" are highlighted.
2. Use the assignments of non-passive associated forms to infer a replacement assignment for the passives.
3. Make a global substitution of all *passive* flags to *intransitive*.
4. Iteratively, convert the intransitive flags to other flags, as indicated by the analysis of the dictionary.

The results of carrying out this program are far from perfect, but they did get rid of the noxious *passive* 'semantics.' A systematic check of the replacement assignments is in order.

[6] Identical vectorings hold for the three senses of –שֶׁ as well as its cousins.