
Hagen Peukert
University of Hamburg
hagen.peukert@uni-hamburg.de

**Proving poverties of the stimulus wrong:
A computational and corpus-based case in point**

The Argument from the Poverty of the Stimulus (APS) still dominates a substantial part of the research in theoretical linguistics. While in generative approaches, the APS provides the foundation of the innateness hypothesis; in other paradigms (usage-based, interactionist, cognitivist) the argument is central because it is argued against it.

Poverty of the stimulus arguments are indeed theoretical because they do little to empirically back their premises. Following Pullum & Scholz (2002) the logic of the argument has two premises. First, language is either genetically encoded or learned data-driven. The second premise, however, is meant to be empirical, that is, infants learn what is not given in the input. Since, by definition, data driven means that everything can be learned from the input, the empirical premise excludes data driven strategies as an option. Now, only one alternative remains in the initial premise. Thus one can deduce modus tollens the genetically encoded strategy as the correct alternative via exclusion. Although the conclusion is correct, the result can be questioned because the premises are far from plausible nowadays. Accepting that the two alternatives postulated in the first premise are somewhat sufficient, the empirical premise seems more problematic. Also known as lack of evidence, this premise is a simple assertion without a systematic and comprehensive proof. The assumption was certainly still acceptable at the time of its conception. Today, however, we have collected specific and representative corpora, developed a body of sophisticated methodologies for quantitative analyses, and used the computational power to carry out complex

calculations on several variables at a time. This development enables us to re-evaluate the argument from the poverty of the stimulus.

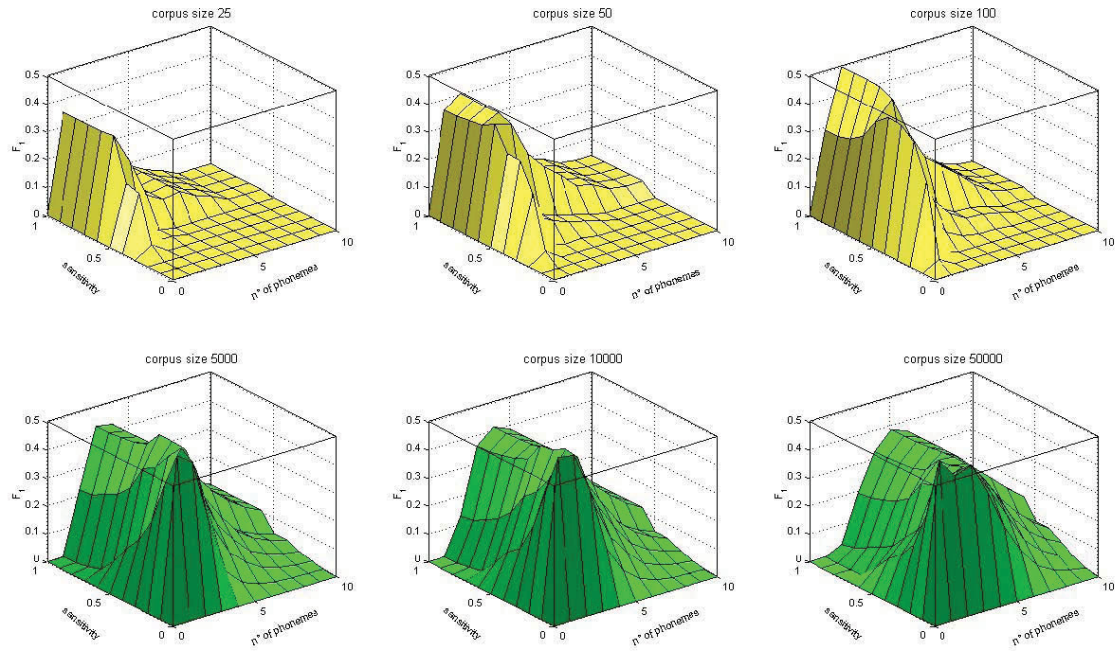
Even though the discussion on the APS centered around the syntactic analysis of phrase structures within the linguistic community (Clark 2003:18), all of the three other cognitive capabilities originally identified by Chomsky (Chomsky 1977:152; see also Dittmann 2006:74) – segmenting the speech stream, identifying word categories and syntactic categories – still deserve equal attention. Up to now, word segmentation belongs to the set of phenomena in language acquisition that cannot be solved since the input is impoverished. As a case in point, for a representative corpus of English child-directed speech, we will show that the speech stream contains enough information, although hidden, so that words and phrases can be reliably located. We model a new algorithm for word segmentation that only includes general cognitive abilities of six to eight months-olds. The model will not hide any language specific information; it is data-driven, bottom-up, usage-based. The guiding question is how do six to eight month-olds segment a stream of sounds into words? Our suggestion is to build a model around the specific distribution of sounds of a language, that is, transitional probabilities. The model at hand thus bridges Phonology and Semantics in the acquisition process.

The constraints defined in this model are all based on general cognitive abilities. First, the most important constraint is the inclusion of transitional probabilities in the model. Second, it is important to pin down the exact form of the unit of perception. From this unit, calculations of transitional probabilities will be accomplished (Saffran et al. 1996). Here the phoneme seemed best suited. Using phonemes allows for a more elaborated inquiry by controlling both possibilities: syllables as well as all other existing combinations of phonemes. Third, it is clear that babies have to memorize words for a longer period of time (Jusczyk & Aslin 1995; Jusczyk & Hohne 1997), so that, fourth, the most frequent ones (Shi et al. 2006; Jusczyk et al. 1994) can be mapped top-down (Bortfeld et al. 2005) onto an unknown speech input.

The last implementation of the model to be presented takes representative samples of a controlled size from CHILDES (MacWhinney 1995), converts them into an IPA-format, deletes all stress information and white spaces, calculates the transitional probabilities for each combination of phoneme chains ranging from 1 to 10 phonemes (n-gram model) and marks white spaces when a threshold value is reached. The threshold value is a variable manipulating the sensitivity of the child and is defined between 0 and 1. Thus, a second loop runs through ten values of a threshold limit and outputs its maximum for each phoneme chain combination.

Once the value at the optimum of both variables (partial derivative of length of phoneme chain and sensitivity of the child), is calculated, the most frequent items (differing between 5 and 30 ‘words’) are selected and saved in a list. Eventually the next corpus is input and processed as described above.

FIGURE 1. SEGMENTATION PERFORMANCE (F1) FOR SPECIFIED PARAMETER COMBINATIONS (LENGTH OF PHONEME CHAIN AND CHILD'S SENSITIVITY) AT GIVEN CORPUS SIZES.



The simulation allows measuring the exact size of the corpus that is necessary to make a maximum of correct segmentations as a function of the child's sensitivity and a corresponding unit of perception (Figure 1). The most important discovery of the simulation is an important property of the segmented corpus. It is indeed true that only about one third of all segmented words are correct (Brent 1999) and that this number cannot possibly account for a starting point in the segmentation process. At first sight it seems convincing because the child cannot know which words are correct segmentations (Gambel/Yang 2005). Searching the corpus for more detail, it becomes clear that the wrongly segmented corpora all encode some more important information, which is not at all obvious looking at it superficially. The most frequent segmentations of the wrongly segmented corpus happen to be lexical items.

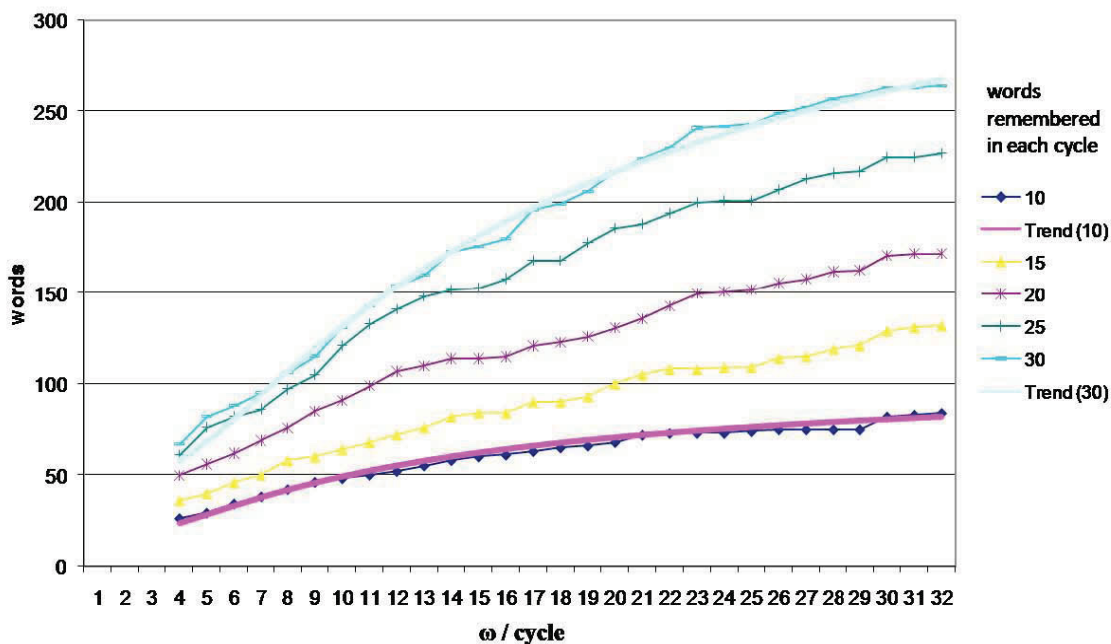
Only that finding allows solving the segmentation problem since out of each representative corpus a certain number of words can now be extracted reliably (Table 1). This number is described by a function and encodes the number of words an infant should be able to memorize (Figure 2).

$$f(\omega) = \frac{12,51\xi - 35,1}{e^{\frac{\xi}{5\omega} + \frac{8}{\omega}}} + \frac{31}{20}\xi + 1,5$$

(for $2 \leq \omega \leq \infty$, $10 \leq \xi \leq 30$, whereas ζ is the function specifying the words to be memorized and ω is the number of corpora or cycles passed without regress to a lexicon)

In addition, the set can serve as 'learning material' for language specific rules (prosody, allophonic variation, phonotactics).

FIGURE 2. GROWTH OF WORDS PER CYCLE (ACTUAL AND IDEAL).



For arguing against the argument from the poverty of the stimulus, it is sufficient to show that the information is actually existent in the input since the APS is itself theoretical and does not provide empirical evidence. The model to be presented goes a step further and shows that infants of the respective age possess the necessary cognitive abilities as well. The next step would be an experimental setting confirming the interrelation and order of the suggested processes.

TABLE 1. PRELEXICAL DEVELOPMENT FOR $\omega = 6$ CYCLES AND $\zeta = 10$ ITEMS REMEMBERED DURING EACH CYCLE.

Rank	$\omega = 1$	$\omega = 2$	$\omega = 3$	$\omega = 4$	$\omega = 5$	$\omega = 6$						
1	ðæt	20	ðæt	34	ju	46	ðæt	58	ðæt	73	ðæt	93
2	wæt	19	ðer	32	ðæt	46	ju	57	ju	71	ju	89
3	jə	19	mɔrgʌn	24	ðer	45	ðer	45	jɔr	47	ænd	56
4	lɪtʌl	14	ju	24	lets	31	ɔrju	37	wæt	46	ðæts	48
5	ju	13	wæt	19	ænd	26	wæt	33	ðer	45	jɔr	47
6	ðer	13	jə	19	ɔrju	26	lets	31	ænd	44	wæt	46
7	ɔrju	13	wer	14	hɪr	25	jɔr	28	hɪr	37	ðer	45
8	ænd	12	lɪtʌl	14	mɔrgʌn	24	ænd	26	ɔrju	37	mɔrgʌn	44
9	dʊŋ	12	ðæts	13	wæt	19	wer	25	lets	31	hɪr	37
10	ðerjuɡəʊ	12	ɔrju	13	jə	19	hɪr	25	ðæts	31	ɔrju	37
11			ænd	12	əʊkeɪ	17	mɔrgʌn	24	wer	25	lets	31
12			dʊŋ	12	wer	14	həv	23	mɔrgʌn	24	wer	25
13			ðerjuɡəʊ	12	lɪtʌl	14	jə	19	həv	23	ŋ	24
14			mɔmi	11	kən	14	əʊkeɪ	17	jə	19	həv	23
15			lets	11	wi	13	red	16	əʊkeɪ	17	jə	19
16			jɔr	11	ðæts	13	lɪtʌl	14	red	16	əʊkeɪ	17
17			hɪr	11	ŋ	12	kən	14	lɪtʌl	14	red	16
18					həv	12	wi	13	kʔn	14	lɪtʌl	14
19					dʊŋ	12	tu	13	wi	13	kən	14
20					ðerjuɡəʊ	12	ðæts	13	tu	13	wi	13
21					mɔmi	11	ŋ	12	kʌdlɪtedɪ	13	tu	13
22					jɔr	11	dʊŋ	12	lʊk	12	pɪrtɪ	13
23							ðerjuɡəʊ	12	kʌm	12	kʌdlɪtedɪ	13
24							mɔmi	11	ŋ	12	wɔtɜ	12
25							ðʌ	11	hænd	12	ʌp	12
26									dʊŋ	12	lʊk	12
27									ðerjuɡəʊ	12	kəm	12
28									mɔmi	11	ɪts	12
29									ðʌ	11	hænd	12
30											dʊŋ	12
31											ðerjuɡəʊ	12
32											ɔr	12
33											mɔmi	11
34											ðʌ	11

References

- Bortfeld, Heather, Morgan, James L. und Golinkoff, Roberta Michnick 2005. Mommy and me. *Psychological Science* 16(4), 298-304.
- Brent, Michael R. 1999. Speech segmentation and word discovery. In *Trends in Cognitive Sciences* 3(8), 294-301.
- Chomsky, Noam. 1977. *Reflections on Language*. New York: Pantheon Books.
- Chomsky, Noam 1988. *Language and problems of knowledge*. Cambridge: MIT Press.
- Clark, Eve V. 2003. *First language acquisition*. Cambridge: Cambridge University Press.
- Dittmann, Jürgen 2006. *Der Spracherwerb des Kindes*. München: Beck.
- Gambell, Timothy and Yang, Charles D. 2005. *Mechanisms and constraints in word segmentation*. ms, Yale University.
- Jusczyk, Peter W. and Aslin, Richard N. 1995. Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology* 29(1), 1-23.
- Jusczyk, Peter W. and Hohne, Elizabeth A. 1997. Infants' memory for spoken words. *Science* 277, 1984-1986.
- Jusczyk, Peter W., Luce, Paul A. und Luce, Jan Charles 1994. Infants sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*. 33(5), 630-645.
- MacWhinney, Brian 1995. *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale: Lawrence Erlbaum.
- Pullum, Geoffrey K. and Scholz, Barbara C. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review* 19(1), 9-50.

- Saffran, Jenny R., Aslin, Richard N. and Newport, Elissa L. 1996. Statistical learning by 8-month-old infants. *Science* 274(5294), 1926-1928.
- Shi, Rushen S., Cutler, Anne, Werker, Janet and Cruickshank, Marisa 2006. Frequency and form as determinants of functor sensitivity in English-acquiring infants. *Journal of the Acoustical Society of America* 119(6), E161-E167.
-