Phonemic Text Transcription Enhances Automated Morpheme Detection: The Importance of Knowing Which Information is Used From the Input

Hagen Peukert[1]

[1]Universität Hamburg, ZFDM, Germany

Abstract

The identification of words in texts and speech is an important ingredient in speech and language recognition systems. Unsupervised learning algorithms use distributional information in texts to derive regularities that the human brain would construe as lexical units, i.e. morphemes. Since statistical distributions of alphabetic or phonemic clusters are not knowingly experienced by the human mind, the exploitation of such information by a machine and making it accessible to our senses, results in new insights of the input material. The study at hand focuses on the properties of language input, which is systematically varied. It will be shown that the language register will not have an effect in English on the identification of words. Yet, there are significant differences if the form of representation is changed from an alphabetic to a phonemic representation. The control language, Japanese, reveals that it is not a universal feature among the languages of the world. Hence the design of algorithms exploiting distributional cues should be defined language-specifically.

Keywords

Morpheme Boundary Detection, Statistical Learning, Phonemic Distribution

1. Introduction

1.1. Motivation

The detection of word boundaries and hence the identification of words is an intensively researched and growing field since speech recognition has been rediscovered as a subject of Artificial Intelligence (AI). For written texts in languages such as Chinese or Japanese, in which words are not explicitly marked by surrounding spaces, the quick detection of words is important for any form of elaborated text processing. Even in English, one can think of scenarios where white spaces as markers of words are misleading (e.g. compounds). The many algorithms developed in Named Entity Recognition illustrate how challenging a problem considered comparatively easy may become. The solutions range from simple table look-ups to the inclusion of complex syntactic structures.

The design of learning algorithms and pattern recognition in general, centers around the question of how much external information is added for fine tuning and how much information is

Humanities-Centred AI (CHAI), Workshop at the 44th German Conference on Artificial Intelligence, September 28, 2021, Berlin, Germany

hagen.peukert@uni-hamburg.de (H. Peukert)
0000-0002-3228-316X (H. Peukert)

CEUR Workshop Proceedings (CEUR-WS.org)

gained from the input. The later is by far the more recognized and elegant way of solving learning problems. It has consequences for the processing efficiency and the cost of involved language resources. It saves the burden of encoding millions of exceptions and individual vocabularies. Finding a universal and slim mechanism stated as a general rule that can be formalized with few assumptions as well as little predefined knowledge carries a high explanatory potential for both research in AI, Linguistics, and Cognitive Sciences. So it seems justified to investigate further into the hidden structures of language, its building blocks and their distribution with the aim of minimizing the use of external information.

1.2. Research Question

For reasons of brevity, from the mass of potential research questions implied by the motivation, two research questions crystallize as more relevant to the problem of word identification. First, to which exact extent do different representations of input material differ in their statistical distribution of the units of perception (i.e. syllables, clusters of letters or phonemes) and how does it affect the segmentation performance? Second, can we treat the differences of transitional probabilities between and within words as a universal fact that is true for all natural languages in general?

A reasonable starting point for the above research questions is to observe real life, that is, to find out how humans master this problem. Given that every newborn is able to learn any of the six to eight thousand languages of the world if exposed to sufficient language input, it is clear that language-specific knowledge such as supra-segmental details (prosodic patterns), phonotactics, or even word knowledge is not present in the minds of the learner. As a short description of the procedure for the project at hand, the approach followed in Cognitive Science – observing language acquisition from real life and carrying out experiments to determine the amount of general strategies, which the language learner starts with – is taken as an input here. As characteristic to the Humanities, we will make theoretic, but plausible, assumptions by describing a concrete mechanism, i.e. a model how word identification could be carried out given the real world abilities of the language learner that were experimentally proven by Cognitive Science.

In figure 1 this process is visualized as a hermeneutic circle, in which Cognitive Science has its core competencies in delivering experimental results [1]. In fact, psychologists do so to contribute to theory building and modeling, however, together with linguists. Linguistic expertise is also needed for programming and carrying out computer simulations. It is there where Artificial Intelligence will come into play once the underlying principles are understood to the extent that enables the researcher to devise working technologies without being based on behavioral experiments.

To illustrate this point, parallel to the process of learning to fly, humans first studied the mechanisms and principles of birds. Experiments soon made us discover principles of aerodynamics. And finally a fly technology was invented that use different mechanisms for flying than birds, but still work according to the same general principles of aerodynamics that also birds are subdued to. Hence in analogy, by studying the mechanisms of speech perception and segmentation at the very early stages of child language development in Cognitive Science, we like to discover the very general principles. Once the principles of speech segmentation are

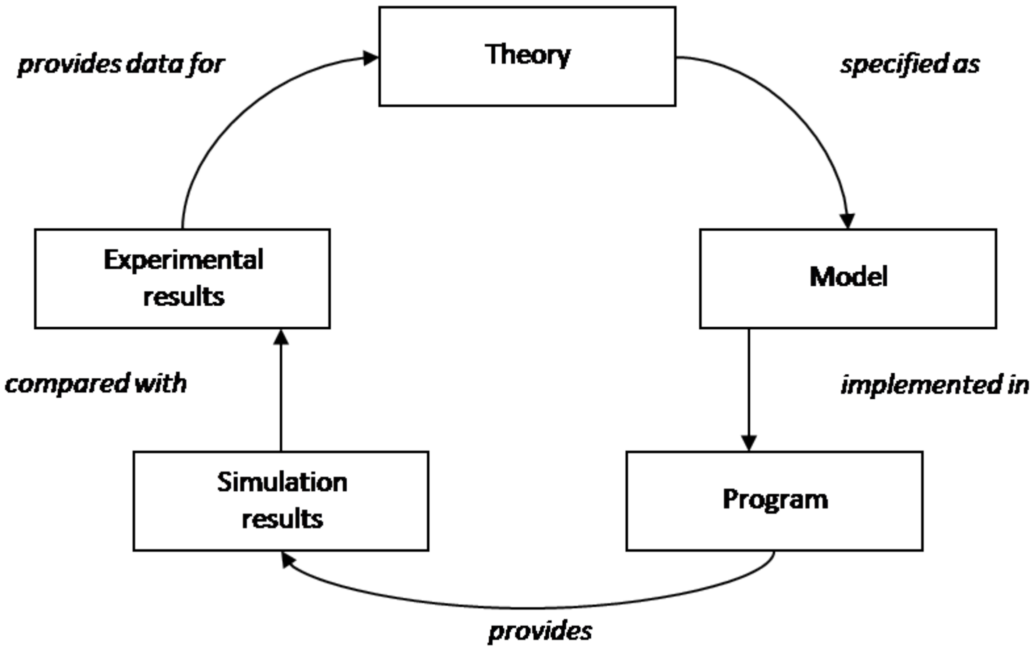

Figure 1: Psycholinguistic Modeling [1]

fully understood, Artificial Intelligence can be exploited to develop technologies beyond human capabilities, but able to have the same or better results in recognizing words in running speech.

2. Background and Review

There are essentially two possible approaches to identify a word: *word-based* and *boundary-based*. The first approach presupposes that a word is given in some kind of either a concrete or more abstract description. The challenge here is to match this word to a chain of words while taking into account phenomena such as the exact allocation or the embeddedness (e.g. *and* in *understand* or *is* in *fist*). The word-based approach is a more direct and intuitive procedure. It simulates the idea of an internal lexicon functioning roughly like a template. When adults learn a foreign language and hear a sound sequence which they hypothesize to be a lexical unit of its own, they often consult a dictionary to search this item. In addition, the adult learner is also interested in the meaning of the identified sequence. Likewise the word-based approach may serve two purposes: firstly, test if the sequence actually exists and, secondly, understand the meaning of a word. A large body of literature was produced, which is not be further considered here. Suffice it to mention that the foundations [2] after refinement and improvement were implemented in well-studied algorithms such as *INCDROP* [3] or *BootLex* [4].

Word boundary-based approaches shifted the focus from the allocation of a possible word candidate to the linguistic environment, in which words occur, i.e. knowing where each word of a sequence begins (or ends) is sufficient to identify each word. Put differently, the beginnings and endings of words define implicitly the entire word even if nothing is known about what is between the boundaries. Boundary-based approaches subdivide into *utterance-boundary* and *predictability* models. The former tend to be rule-based and they are derived from systematic observations of linguistic regularities.

A lot of attention has been paid to supra-segmentalia of languages in the hope to find a general principle which unites the world's languages. Although a general principle is not in sight, research in this area has been fruitful as the example of metrical segmentation strategies illustrates [5, 6, 7, 8]. Using phonotactics – sound combinations that do not at all occur together – hint at at the beginning or endings of words (e.g. lr or the velar nasal ŋ never shows up word-initially in English or German) for obvious reasons [9]. Finally, allophonic cues [10, 11, 12] use the systematic variation of sounds dependent on their position in the word (word-initial, word-final) to determine a word boundary. As an example in English, consider the difference between the velarised variant [ɫ] of the alveolar lateral approximant [l] in feel versus love or the aspiration of word-initial stops such as p^{h} in pulpit.

Predictability models as the other subgroup of boundary-based approaches can best be subsumed under the notion of *Statistical Learning*. Essentially this family of models looks at the distribution of whatever the unit of perception (syllables, diphones, phonemes, letters) is and assimilates the observation that sound clusters within the word are significantly different than across words. This can theoretically be shown in text corpora [13, 14, 15] and practically in experiments [16]. While the idea is not new [17, 18], the prove that human learners – particularly adults and 6-months-olds – make extensive use of the mechanism gained momentum just two decades ago [19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35]. Unlike any of the other approaches statistical learning does not assume additional materials to learn from but the input text. This is different for word-based approaches, for which the entire lexicon has to be available. Metrical segmentation, allophonic cues, or phonotactics also presume an initial, though small, list of words containing candidates, which reveal the language-specific properties that should be learned.

In the search of a plausible explanation, where such material could stem from, the argument of isolated words [6, 36] as a starting point for deriving all kinds of language specific features that could subsequently be used for segmenting unknown speech input has received some recognition in the literature. Once the learner had a list of words, she or he would be able to derive prosodic patterns, allophonic variation, or phonotactic pecularities of the language under investigation. And with this knowledge a mechanism could be derived allowing to segment any running speech into words. To assume a list of isolated words, shifts the problem to an earlier stage of language learning because the question remains unchanged: how can the list be produced? The essential riddle to be solved stays as is: without prior knowledge how can a language learner figure out what a word is and what are its parts or even larger units if the only input given is the speech stream? It is a robust effect from various experiments and corpus studies that words hardly ever occur in isolation. Even if mothers are explicitly asked to do so when communicating with their babies, only about 20% of the words actually are isolated [11]. Even if accepting this, however, it would not help much since the little toddler has no chance to know when a word occurs in isolation. Also the suggestion that some kind of general word analyzer is part of the genetic endowment of humans cannot hold since there are simply too many languages that show to much variation that even hypothesizing and testing sequences of sounds could never converge to the same system of words in the observed amount of time. This dilemma is the *chicken and egg* problem of word segmentation. Without prior access to words, how can a learner know what a word is?

The answer to this question is to apply predictability strategies. They make no language-specific presumptions. Statistical learning assumes a general cognitive ability to have mental representations of frequency counts and relations of frequenciey to each other. On a more abstract level, these capabilities are also needed for other mechanisms in our cognitive apparatus, e.g. the visual systems builds equally complex patterns and computations around conditional probabilities of light signals.

3. Methodology

To answer these research questions a model is applied that has been tested and discussed previously [37],[1] but it is used here to systematically vary the input of phonemically transcribed and alphabetic texts of different languages and its registers. The model at hand can be described as a mixture from all the approaches elaborated on above. First, the model follows strictly the boundary-based predictability approach to arrive at a word list whose content will be exploited by a word-based approach to aline the words to new input. Indeed, it would be possible to extract additional linguistic features as suggested by the utterance-boundary approaches. However, this will only serve secondary purposes with regard to the research questions. So it is left for follow-up research in the future.

Put briefly, the algorithm used here can best be described in terms of an n-gram model calculating the optimal segmentation results based on the transitional probabilities for all n-gram combinations of a given input. Here the predictability approach is applied. The input are texts in alphabetic and IPA transcribed form, from which all white spaces have been deleted. The model also incorporates the manipulation of the length of phoneme chains or a lexicon function. The lexicon function allows to save e.g. the 10 or so most frequent segmentations of the input in an extensible list, which can be used as an additional source of information in the segmentation process. This is where the word-based models are taken into account.

Three different text corpus collections are taken as an input: samples from Child Language Data Exchange System (CHILDES) [39], International Corpus of English – Great Britain (ICE-GB), and Corpus of Spoken Japanese (CSJ) [40]. CHILDES is used because it is clear from the literature that child-directed speech (CDS) is beneficial to the learning of the language for its repetitions and, allegedly, for the peculiar distributions of the units of perception [41]. Japanese is selected because the Kanji writing system does not encode word boundaries. So readers and speakers alike might rely on the same kind of cues. It would be enlightening to see if these cues are based on transitional probabilities as well in addition to the known rhythmic (the morae), lexical, and syntactic information.

[1] There is a tool set [38] that can be used to run the simulations whose results are presented below

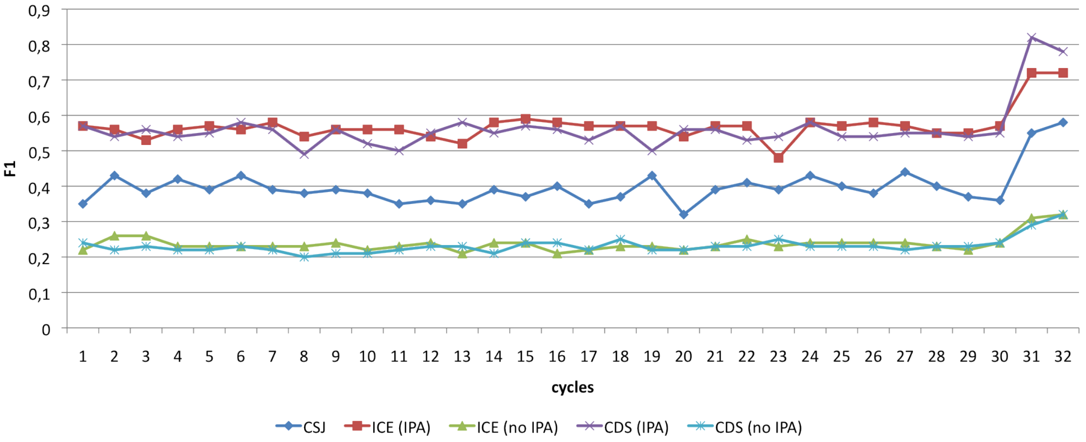

Figure 2: F1-measures for CSJ, ICE, and CDS (in phonetic and alphabetic script)

4. Results

The results are graphically visualized in figure 2. The ordinate labeled *cycles* encodes the number of different texts that were input, that is, either in phonemic (IPA) or alphabetic (no IPA) transcription. At each cycle the most frequently segmented clusters are memorized as a lexicon and this is used in cycle 30 to boost the segmentation performance to an acceptable level. The output is given as an F1-measure, which includes precision and recall at equal proportions. With respect to the first research question, the two English registers, ICE and CDS, behave by and large equally. While both, the alphabetic representations for *ICE (no IPA)* and *CDS (no IPA)*, show low F1 levels at around 0.3, which is insignificantly different from a random segmentation, their phonemic counterparts (labeled as *ICE* and *CDS*) reveal equal rates of improvement. This suggests that the distribution of phonemes is more supportive for an segmentation strategy exploiting the transitional probabilities between and within words independent of the language register. Also, the English registers depict similar growth rates when the generated lexicon is used.

The second research question can be answered if the data on Japanese *(CSJ)* is considered. Although beyond randomness especially after the usage of the lexicon, the phonemic representation of Japanese cannot reach an acceptable level of performance. The Japanese segmentation performance is somewhat between the performance of the English phonemic and alphabetic representations. So it is unlikely that Japanese contains the same distributional pattern as English that is useful for segmentation. And this suffices for the claim that predictable phoneme clusters based on transitional probabilities to detect word boundaries cannot be a language universal phenomenon.

5. Discussion

Since English CDS is predictable on the basis of typical phoneme clusters from English adult speech if logistic regression is applied [42], we know that English adult speech differs significantly from English CDS in its distribution of sound chains. However this difference has no significant effect on the segmentation performance. And thus a different distribution does not necessarily account for any prediction of the segmentation performance. It is now plausible to assume that one can make the same claim for Japanese following the assumption that the segmentation process as such could be a general cognitive ability initiated by statistical learning of the input. The study revealed that this is not the case. Although it holds for the qualities of English, it does not for Japanese.

This is interesting for two reasons. First, on the one hand, the English case illustrates that the natural grown system of actual sound representations and its distribution contains information useful for the segmentation process. There is ample evidence that the optimization principle of language has been pushing the order of phonemes and syllables according to the needs of the environment and the capacity of the brain in the present direction that eases the recognition of lexical units.

The artificial system of language representation, that is, the writing system that has culturally grown, misses this information clearly because distributions are hidden and the inventors of the script surely could not have thought of it. On the other hand, the Japanese case makes clear that even natural languages develop different segmentation strategies that do not necessarily rely on transitional probabilities and that, in addition, natural languages do not necessarily optimize in the same direction. All cases taken together suggest that the mechanisms of language change might not have achieved the optimal distribution of phonemes yet.

Second, for researchers of AI the results show that the details of the input cannot be ignored independent from the size of the data to be learned from. In the case of English corpora, the above result – the form of representation and not the language register is crucial – was not to be expected since the two forms of text representation are uniquely transferable in one another and hence it seemed plausible that their distribution would do too. This assumption did not prove right.

Now the idea for further investigation is to find out whether there is an alphabet that also fulfills the criterion of being uniquely transferable back and forth to all English script systems, but is further optimal with regard to its segmentation performance or even allows perfect segmentation. This is the case when the phoneme clusters within a word are nearly always different than across words. Moreover, the question arises whether such an ideal text representation could be artificially constructed for Japanese or other languages. If so, AI could once again compensate the shortcomings of natural sound pattern optimization.

References

[1] T. K. d. S. Dijkstra, Computational Psycholinguistics, Taylor & Francis, London, 1996.

[2] D. C. Olivier, Stochastic grammars and language acquisition mechanisms, Harvard Universität, Cambridge, 1968.

[3] M. R. Brent, An efficient, probabilistically sound algorithm for segmentation and word discovery, Machine Learning 34 (1999) 71–105.

[4] E. O. Batchelder, Bootstrapping the lexicon: A computational model of infant speech segmentation, Cognition 83 (2002) 167–206.

[5] C. H. Echols, M. J. Crowhurst, J. B. Childers, The perception of rhythmic units in speech by infants and adults, Journal of memory and language 36 (1997) 202–225.

[6] E. K. Johnson, P. W. Jusczyk, Word segmentation by 8-month-olds: When speech cues count more than statistics, Journal of Memory and Language 44 (2001) 548–567.

[7] J. L. Morgan, Extracting sentence structure from infant-directed speech, Infant Behavior & Development (1996) 632.

[8] P. W. Jusczyk, A. Cutler, N. J. Redanz, Infants' preference for the predominant stress patterns of english words, Child development 64 (1993) 675–687.

[9] A. D. Friederici, J. M. I. Wessels, Phonotactic knowledge of word boundaries and its use in infant speech perception, Perception & Psychophysics 54 (1993) 287–295.

[10] P. W. Jusczyk, How infants begin to extract words from speech, Trends in Cognitive Sciences 3 (1999) 323–328.

[11] R. N. Aslin, J. Woodward, N. LaMendola, T. Bever, Models of word segementation in fluent maternal speech to infants, in: J. Morgan, K. Demuth (Eds.), Signal to Syntax: Bootstrapping From Speech to Grammar in Early Acquisition, Erlbaum, Providence, 1996, pp. 117–134.

[12] J. A. Christiansen, M. S. Seidenberg, Learning to segment speech using multiple cues: A connectionist model, Language and Cognitive Processes 13 (1998) 221–268.

[13] Z. S. Harris, Morpheme Boundaries Within Words: Report on a Computer Test, Transformations and Discourse Analysis Papers 73, volume 1 of *Papers in structural and transformational Linguistics*, University of Pennsylvania, Philadelphia, 1967.

[14] D. E. Rumelhart, J. L. McClelland, Pdp models and general issues in cognitive science, in: J. L. McClelland, D. E. Rumelhart (Eds.), Parallel Distributed Processing, volume 1, MIT, Cambridge, 1986, pp. 110–149.

[15] J. L. Elman, Finding structure in time, Cognitive science 14 (1990) 179–211.

[16] J. R. Saffran, R. N. Aslin, E. L. Newport, Statistical learning by 8-month-old infants, Science 274 (1996) 1926–1928.

[17] Z. S. Harris, From phoneme to morpheme, Language 31 (1955) 190–222.

[18] Z. S. Harris, Distributional structure, Word 10 (1954) 146–162.

[19] P. Cairns, R. Shillcock, N. Chater, J. Levy, Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation, Cognitive Psychology 33 (1997) 111–153.

[20] R. N. Aslin, J. R. Saffran, E. L. Newport, Statistical learning in linguistic and nonlinguistic domains, in: B. MacWhinney (Ed.), The Emergence of Language, Erlbaum, London, 1999, pp. 359–380.

[21] R. N. Aslin, J. R. Saffran, E. L. Newport, Computation of conditional probability statistics by 8-month-old infants, Psychological Science 9 (1998) 321–324.

[22] J. L. Morgan, J. R. Saffran, Emerging integration of sequential and suprasegmental information in preverbal speech segmentation, Child Development 66 (1995) 911–936.

[23] J. R. Saffran, Words in a sea of sounds: the output of infant statistical learning, Cognition 81 (2001) 149–169.

[24] J. R. Saffran, The use of predictive dependencies in language learning, Journal of Memory and Language 44 (2001) 493–515.

[25] J. R. Saffran, Constraints on statistical language learning, Journal of Memory and Language 47 (2002) 172–196.

[26] J. R. Saffran, Birds do it - why not babies?, Developmental Science 6 (2003) 46–47.

[27] J. R. Saffran, Statistical language learning: Mechanisms and constraints, Current Directions in Psychological Science 12 (2003) 110–114.

[28] J. R. Saffran, Absolute pitch in infancy and adulthood: the role of tonal structure, Developmental Science 6 (2003) 35–43.

[29] J. R. Saffran, D. P. Wilson, From syllables to syntax: Multilevel statistical learning by 12-month-old infants, Infancy 4 (2003) 273–284.

[30] J. R. Saffran, J. F. Werker, L. A. Werker, The infant's auditory world: Hearing speech, and the beginnings of language, in: R. Siegler, D. Kuhn (Eds.), Handbook of Child Development, John Wiley & Sons, New York, 2006, pp. 58–108.

[31] J. R. Saffran, E. L. Newport, R. N. Aslin, R. A. Tunick, S. Barrueco, Incidental language learning: Listening (and learning) out of the corner of your ear, Psychological Science 8 (1997) 101–105.

[32] J. R. Saffran, E. K. Johnson, R. N. Aslin, E. L. Newport, Statistical learning of tone sequences by human infants and adults, Cognition 70 (1999) 27–52.

[33] J. R. Saffran, K. Reeck, A. Niebuhr, D. Wilson, Changing the tune: the structure of the input affects infants' use of absolute and relative pitch, Developmental Science 8 (2005) 1–7.

[34] J. R. Saffran, A. Senghas, J. C. Trueswell, The acquisition of language by children, Proceedings of the National Academy of Sciences of the United States of America 98 (2001) 12874–12875.

[35] J. R. Saffran, E. D. Thiessen, Pattern induction by infant language learners, Developmental Psychology 39 (2003) 484–494.

[36] M. R. Brent, J. M. Siskind, The role of exposure to isolated words in early vocabulary development, Cognition 81 (2001) B33–B44.

[37] H. Peukert, Kindliche Kalkulationen: Eine Computersimulation über den Einfluss stochastischer Informationen auf die Wortsegmentierung beim Erstspracherwerb, Kassel University Press, Kassel, 2009.

[38] ZFDM, Gitlab Repository Universität Hamburg, 2021. URL: https://gitlab.rrz.uni-hamburg.de/softwaretools/segmentation.

[39] B. MacWhinney, The CHILDES project, Erlbaum, Hillsdale, 1995.

[40] S. Furui, K. Maekava, H. Isahara, The corpus of spontaneous japanese, 2004. URL: https://ccd.ninjal.ac.jp/csj/en/index.html.

[41] R. N. Aslin, J. Woodward, N. LaMendola, T. Bever, Models of word segementation in fluent maternal speech to infants, in: J. Morgan, K. Demuth (Eds.), Signal to Syntax: Bootstrapping From Speech to Grammar in Early Acquisition, Erlbaum, Providence, 1996, pp. 117–134.

[42] H. Peukert, Hidden structures in english corpora, in: D. Mukherjee, M. Huber (Eds.), Corpus Linguistics and Variation in English: Theory and Description, volume 75 of *Language and Computers*, Rodopi, Amsterdam, 2012, pp. 131–141.