Figure 5: Lower left: abstract from Japanese page of Figure 1; note Romaji in 2nd and 8th line do not follow same pitch, and compressed 6th line. Upper: normalized vertical projection. Right: magnitude of 1024-point DFT of normalized signal; peak corresponds to 9.2 symbols/inch (peak contains 92% of energy in squared magnitude, including secondary peak at 18.4 symbols/inch).

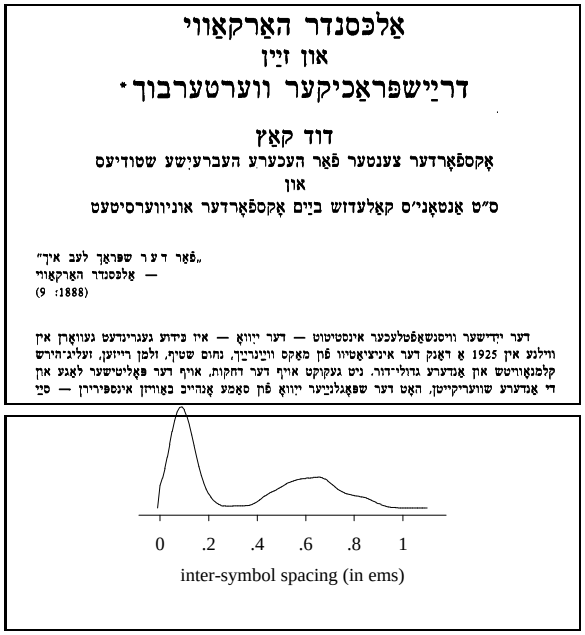

Figure 6: Upper: Yiddish text. Lower: smoothed histogram of inter-symbol spacing (scaled by square root), bimodal distribution is clear despite different text sizes ($t = .28$ ems); all words were isolated properly then ordered right-to-left within line.

in the range [0.1,0.4] ems for all the languages we have studied.

The threshold t is then applied to each line in turn, scaled by text size in the line, to distinguish word spaces from inter-symbol spaces. The resulting words are sorted into the reading order of the language. Figure 6 illustrates this method when applied to Yiddish text (Yiddish is a dialect of German written using the Hebrew alphabet).

9 Discussion

We have described a largely language-free approach to the analysis of page layouts into text blocks, lines, words, and symbols. The wide variety of techniques we have used should not conceal a unifying strategy, in which a sequence of model-refinement steps carefully chosen to make available the widest possible statistical support for the decisions to be made at each step. We call this a *global-to-local* strategy. For example, our *first* step is to estimate skew because skew is the only layout parameter for which the statistical evidence is distributed over the *entire* page, and therefore able to be estimated with the highest confidence. There are many other examples of this principle throughout the system. We feel that this statistically conservative strategy helps explain the system's versatility in spite of the fact that it does not rely on backtracking control or detailed application-dependent layout models.

We have also encountered software engineering challenges in building a strongly language-free system. For example, the implementation of all algorithms following the inference of text line orientation have been generalized to cope with horizontal and/or vertical lines. Also, all algorithms which deal with symbols or groups of symbols, such as shape-directed resegmentation and contextual analysis, have been generalized to handle all possible reading orders of text lines and symbols.

Acknowledgements

We gratefully acknowledge the contributions of Steve Fortune (Delauney triangulation, empty rectangle enumeration) and Susan Dorward (greedy white covers). We benefited from stimulating discussions with Tin Kam Ho.

References

[1] H. S. Baird, "The Skew Angle of Printed Documents," *Proc., 1987 Conf. of the Society of Photographic Scientists and Engineers*, Rochester, New York, May, 1987.

[2] H. S. Baird, "Global-to-Local Layout Analysis," *Proc., IAPR Workshop on Syntactic and Structural Pattern Recognition,* Pont-à-Mousson, France, September, 1988.

[3] H. S. Baird, S. E. Jones, S. J. Fortune, "Image Segmentation by Shape-Directed Covers," *Proc., IAPR 10th Int'l Conf. on Pattern Recognition*, Atlantic City, NJ, June, 1990.

[4] H. S. Baird, "Anatomy of a Versatile Page Reader," *IEEE Proceedings*, July, 1992.

[5] H. S. Baird, "Background Structure in Document Images," *Proc., IAPR Workshop on Structural and Syntactic Pattern Recognition (SSPR'92)*, Bern, Switzerland, August, 1992.

[6] *The Chicago Manual of Style*, 13th Edition, The University of Chicago Press, Chicago, Illinois, pp. 696-701, 1982.

[7] F. Coulmas, *The Writing Systems of the World*, Basil Blackwell: Oxford, 1989.

[8] D. J. Ittner, "Automatic Inference of Textline Orientation," *Proc., Symposium on Document Analysis and Information Retrieval,* Las Vegas, Nevada, April 1993.

[9] S. Kahan, H. S. Baird, T. Pavlidis, "On the Recognition of Printed Characters of any Font or Size," *IEEE Trans. PAMI*, March, 1987.

[10] A. Nakanishi, *Writing Systems of the World*, Charles E. Tuttle: Rutland, Vermont, 1990.