LSM User's Guide

The Formula

$LSM_{preps} = 1 - ((\,|\, preps_1 - preps_2 \,|\,) / (preps_1 + preps_2 + .0001))$

Where $preps_1$ is the percentage of total words in text 1 that were prepositions and $preps_2$ is the percentage of prepositions in text 2.

You calculate this word-level LSM score for each of the following nine LIWC categories:

- personal pronouns (I, their)
- impersonal pronouns (it, those)
- articles (a, the)
- common adverbs (seriously, so)
- auxiliary verbs (am, can)
- conjunctions (but, because)
- prepositions (about, in)
- negations (never, no)
- quantifiers (tons, fewer)

And then finally you average these 9 word-level LSM scores to yield a composite measure of the degree of function word similarity between two texts.

Main Uses

You can use LSM to compare the similarity of any two texts in terms of word frequencies (LIWC output) in order to assess

- how synchronized two people are over the course of an email or postal correspondence
- how in sync two people in a conversation (spoken or computer mediated) are, both on a turn-by-turn and aggregate conversation-level basis
- how in sync a group of people in a conversation are (e.g., in a meeting or at a dinner party)
- how reliably one person uses function words across several different letters or conversations

Function vs. Content Matching

You can also calculate LSM for content word (e.g., nouns, verbs) categories like *negemo* (negative emotions), *posemo* (positive emotions), and all the related emotional categories.

For most purposes, we prefer to use function word synchrony because it's more relevant to the question of whether two people in any given situation are thinking in similar ways. Content word synchrony – especially in situations where content is bound to be similar due to conversation conventions or experimental instructions is often the product of

situational constraints and is unrelated to synchrony or asynchrony in the mental processes of two people.

Content word matching would, however, be extremely interesting in settings where content is more open-ended. So for example, you could look at content synchrony in responses to the picture story exercise (PSE) or the thematic apperception test (TAT) to discover whether two people's PSE results are more similar after playing a Wii™ tennis match against each other than they were at baseline. You could look at power motives, for example.

Punctuation matching might also be really interesting, and it's something that, to my knowledge, hasn't been studied extensively (and never with LSM).

Warning: If you calculate content word matching, I don't recommend averaging cognitive mechanism categories' LSM with the overall function word LSM score. A lot of prepositions and conjunctions are also in the cognitive mechanism categories (e.g., "and" is in both the conjunction and inclusiveness categories). Try to avoid averaging categories with significant overlap. It would make the statistics harder or impossible to interpret.

Using the same logic, don't average LSM from subcategories with LSM for an overarching category. So, for example, don't average the LSM score for first person singular pronouns with LSM for personal pronouns or pronouns in general. It would be redundant and artificially inflate reliability and LSM means.

LSM vs. Correlations

Correlations would look at whether, over several texts or turns, people's language use rises and falls together, or whether changes in one person's language use are associated with changes in the other's language. That's clearly one kind of synchrony. However, correlations don't tell you anything about the discrepancy between the two people's language use. So you could have two people, let's call them Buffy and Angel, in an e-mail correspondence with LIWC results per text as follows:

	Quantifier %	
E-mail pair	Buffy	Angel
1	1.8	6.8
2	3.7	8.7
3	2.9	7.9
4	2.8	7.8
5	1.9	6.9
6	1.2	6.2
7	4.3	9.3
8	4.8	9.8
9	0.7	5.7
10	0.9	5.9

As you can see, Angel always uses 5% more quantifiers than Buffy. That's a pretty large difference for quantifiers. Buffy's mean percent, 2.5%, is almost exactly what the average quantifier use is over thousands of different spoken and written language samples. Angel's mean is 7.5%. So Angel is using about 300% more quantifiers than Buffy and the average person—a huge difference, which LSM reflects. For quantifiers in these ten e-mails, LSM = .46, which is extremely low. The correlation coefficient, however, is a perfect $r = 1.0$. Using only correlations you'd have no way of knowing that quantities are much more salient for Angel than Buffy. Unless you looked at the raw data, of course.

In the end, whether you use correlations or LSM really depends on your research question. If you're looking at verbal mimicry, you really want to know how similar people are *and* whether fluctuations in one's language correspond to changes in their partner's, so LSM is more appropriate.

Interpreting LSM

Here's a table with the expected range for different media and contexts (all values are approximate and based on my memory of past as-yet-unpublished studies):

Context	Min	Max	SD
IM chatting	.80	.90	.05
Correspondence	.85	.95	.07
Poetry	.60	.80	.10
Spoken conversation	.85	.95	.07
Online writing assignments	.65	.80	.07

As you can see, there is more variance for poetry. That's probably because poems aren't written to a person in response to another's poetry like with conversation or correspondence. That said, meaningful (i.e., indicates the same kinds of harmony and disharmony we see in conversations) LSM can be gleaned from professional work such as poetry.

Citing LSM

If you want to use LSM, that's wonderful! I would like to talk with you about your data. If you would like to cite the LSM method in a paper, here's the reference for the original article:

Niederhoffer, K. G. & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology, 21*, 337-360.

And here is the first paper accepted for publication using the updated LSM formula:

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (in press). Language indicators of social dynamics in small groups. *Communications Research*.