Social Variability and Probabilistic Language Processing

Benjamin K. Bergen (bergen@hawaii.edu)
Department of Linguistics, 1890 East-West Rd., Moore Hall 569
Honolulu, HI 96822 USA

Abstract

This paper investigates a particular morpho-phonological process, called French *liaison,* and demonstrates that statistical influences on it from various sources are used in language processing. First, a corpus study shows that liaison is probabilistically affected by phonological, morphological, syntactic, and social factors, some of which interact in their effects on the use of liaison. Next, a perception study demonstrates that language users are able to access these subtle statistical correlations in performing a time-critical categorization task. These results suggest that the acquisition of fluent linguistic knowledge involves at least in part encoding the conditional probabilities of morpho-phonological processes, given social and other independent factors.

Introduction

Long a dirty word among linguists concerned with the mental representation of language, *probability* has in recent years been revitalized as a topic of empirical and theoretical study. A growing body of research (e.g. the papers in Bod, Hay and, Jannedy (2002)) has begun to address the role of probability in the acquisition (Saffran, 2001), processing (Jurafsky, 2002), and representation (Pierrehumbert, 2001) of linguistic units. As the evidence mounts, it becomes progressively more evident that probability is as useful to language learners as they acquire language and language users as they put language to use as it is to language modelers, as they represent language data.

While mentalist linguists have eschewed probability for many years, one class of quantitative sociolinguists has not. The study of socially correlated linguistic variation has for twenty-five years made use of a model called VARBRUL (Sankoff, 1987), which implements a variety of logistic regression. To summarize one of the major results of research of this sort, it seems that in a language community, there is a great deal of linguistic variation, such as the pronunciation or not of a final /r/ in words like *car*. Part of this variation is probabilistically conditioned by social attributes of the speaker such as their social class, or the current speech style (Labov, 1966), among other factors. For example, in Labov's study, middle class speakers were about twice as likely to pronounce final /r/ as were lower class speakers.

With the new focus on probability in the mind of the individual, we are led to wonder whether the statistics of sociolinguistic variation are observed by individual language users, and encoded in their internal language systems. This is the question that the current paper attempts to address:

Are the correlational statistics of sociolinguistic variation in the community reflected in probabilistic language processing in the individual?

French Liaison

I address this question through a set of empirical studies of the phenomenon of *liaison* in French (Tranel, 1981). Liaison is the variable production of a specific set of word-final consonants. Not all final consonants in French are liaison consonants – liaison consonants are specific to particular words or morphemes. Liaison consonants can be the final consonants of words, like *jamais* 'never', or morphemes, like the plural suffix *-s*. These liaison segments can be produced or not, depending on a number of factors.

The strongest of these factors is the quality of the following segment – that is, the first segment of the following word. Liaison consonants are most likely to be produced when followed by a vowel, and are almost never produced when followed by a consonant, or when they fall at the end of an utterance.

For example, the word-final liaison consonant *-s* 'PLURAL' is produced when followed by a vowel, like the final *–s* of the word *grands* in *grands italiens* 'tall Italians'. It is not pronounced, however, when it precedes a consonant, as in *grands russes* 'tall Russians', or when it is utterance-final, as in *les grands* 'the big ones'.

While liaison consonants followed by other consonants are effectively never produced, those which are followed by a vowel (like *grands italiens*) are less predictable. The realization of these pre-vocalic liaison consonants is probabilistic. If all we have to work with is the phonological nature of the following segment, the best we can do is to assign some probability to the realization of a liaison consonant. Before a vowel, that probability is about 0.5.

The following segment, though, is only one of a number of factors that influence liaison. Among these is the grammatical class of the liaison word. While nominal modifiers such as determiners and adjectives favor liaison, most nouns disprefer it. For example, in *des Anglais* 'some English (people)', the final *–s* of *des* 'some' is pronounced, while the final *–s* of the near-homonym noun *dès* 'dice' in a similar context is much less likely to be pronounced, as in *dès anglais* 'English dice'.

Depending on the account, there are ten to twenty other variables that correlate with liaison use (Ashby, 1981; Encrevé, 1988). Some of these, like the identity of the liaison consonant (e.g. /t/ or /r/) or the frequency of the word with the liaison consonant, have been shown to have significant effects (Encrevé, 1988). Most relevant to the current study, three social attributes of the speaker have also

been shown to be significant: age, gender, and level of highest education (Ashby, 1981).

A Corpus Study

The goal in studying French liaison is to investigate the extent to which individuals have internalized non-categorical correlations between phonology and social factors. The perception experiment described in the next section test this knowledge. But in order to evaluate the internalization of probabilistic patterns, we need first to establish what those patterns are in the language of the community – the ambient language from which correlations are extracted. To study variation in a language community, in particular with the exceedingly large number of factors identified in the literature, a large, well-balanced, and controlled corpus is of the utmost importance (Biber, Conrad, & Reppen, 1998).

Data selection

To this end, I based my study on the IDIAP French Polyphone corpus (Chollet et al. 1996). The French Polyphone corpus consists of speech from several thousand French-speaking inhabitants of Switzerland. The speech is of two types. The majority of the corpus is composed of read speech, intended to maximize the range of phonological contexts recorded. A much smaller portion consists of spontaneous human-to-computer speech. All recordings were made over the telephone, meaning that their resolution is somewhat diminished and that there is often background noise. All data is in two forms – the acoustic signal and a human transcription.

I chose to work exclusively on the read speech, for two reasons. First, significantly more data per subject was available in that form. Second, since linguistic interactions between humans and computers remain extremely infrequent, natural human-to-computer speech is extremely variable in its register and style. Including this speech would have added additional factors to my analysis. Since it would be very difficult to evaluate which style or register a particular speaker chose spontaneously to use in a non-circular manner, I chose to restrict analysis of the corpus to read speech.

I randomly selected 200 speakers from this corpus, of whom half were men and half women. Of the 200 speakers I selected, I pruned away those who self-reported as native speakers of a language other than French, since L2 French speakers can be expected to display quite different liaison behavior from natives. This paring down left a total of 173 speakers. Of these speakers, 90 were male and 83 female.

Tagging

I began by identifying all tokens of potential liaison consonants. Some of this was automated - a small set of frequent words, such as *les* 'the' and *dans* 'in' was automatically identified. But, as noted above, not all final consonants are liaison consonants. For example, while the final "s" in *fils* 'strings' is a liaison consonant, pronounceable as either /z/ or ø, the terminal "s" in the homograph *fils* 'son' is always pronounced, as /s/. Because of cases like this, much of the tagging had to be done manually.

Table 1: Variables coded in the Liaison Corpus.

	Variables coded
	Dependent variable: Whether the consonant was produced
1	Orthography of liaison segment, e.g. "r", "s"
2	Phonological realization of liaison segment, e.g. /t/, /z/
3	Type of the preceding segment, e.g. consonant, vowel
4	Orthography of the following segment, e.g. "a", "t"
5	Length of pause between the words
6	Punctuation between two words, as written in the text speakers read from
7	Plural marking of the liaison
8	The grammatical person expressed by the liaison, e.g. 1st, 3rd
9	Grammatical class of the liaison word, e.g. Noun, Preposition
10	Grammatical class of the following word, e.g. Noun, Preposition
11	Frequency of the liaison word in the ARTFL corpus
12	Frequency of the liaison word in the ARTFL corpus
13	Orthographic length of the liaison word
14	Orthographic length of the next word
15	Number of syllables in the liaison word
16	Number of syllables in the following word
17	Speaker's sex
18	Speaker's age
19	Speaker's level of highest education

Once I had found all instances of liaison consonants in the corpus, I turned to factors influencing liaison. On the basis of claims found in the literature (summarized in Bergen, 2001), each potential liaison was analyzed for 19 different factors. Some of these factors were coded automatically, others entirely by hand.

A synopsis of all the dimensions along which each instance of liaison in the corpus was coded can be found in Table 1. Of particular relevance to the following discussion are the social factors: age (young, middle, old), sex (male, female), and level of highest education (primary, secondary, tertiary). Notable also for future reference is factor 9, the grammatical class or pat of speech of the word with a liaison consonant in it.

Analysis

Once the data were coded, they were subjected to statistical analysis. Conditional probabilities in sociolinguistic variation have been principally studied using an analytical tool known variably as VARBRUL (Sankoff, 1987) or GOLDVARB (Rand & Sankoff, ms.). These computer programs implement, among other data collection and preparation tools, a logistic regression algorithm for analyzing the significance and degree of an arbitrarily large number of independent factors on a single dependent linguistic variable.

Of particular interest in the current study is not only whether individuals have internalized probabilistic effects of social factors on linguistic variability. Also relevant are interactions between these factors. The logistic regression included in VARBRUL, though, and consequently in all statistical linguistic studies that make use of this program, is limited to independent factors which do not interact. And yet, there is a natural way to extend the statistical tools used in VARBRUL to cover interactions between factors as well. Logistic regression outside of VARBRUL can include not only terms describing variables multiplied by some constant, but also terms including the product of two independent variables and a constant. Because interactions between factors are of interest at present, full-scale logistic regression, including interaction terms, forms the basis for the statistical analysis of the corpus.

Now that the current study's empirical basis and analysis method have been outlined, we can move on to the crux of this section, the statistical analysis of factors on liaison. The reader will recall that in order to be able to test language users' knowledge of social correlates of probabilistic linguistic effects, we must first establishing what factors probabilistically affect liaison use in the language of the community. In the remainder of this section, we will test both autonomous effects and interaction effects.

Results

In a first statistical test, I ran a step-forward logistic regression on all of the 19 factors enumerated in Table 1 above, without including the possibility of interactions between them. The effects of these factors were tested only as they independently influenced liaison production.

The regression model that resulted from the selection process was able to correctly predict 88.3% of the data. That is, of 2559 tokens, 299 were predicted by the model to have the wrong liaison valence. This is an average to high degree of reliability for a problem with a similarly large problem space.

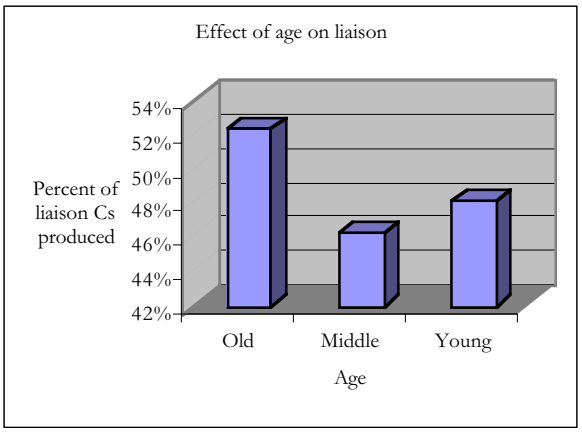
Of the 19 factors originally included, the selection procedure selected 12 for inclusion in the model on the basis of statistical increases of model log likelihood, a statistical metric for whether including a term in the model makes a statistically significant improvement in the model's predictive power. The 12 factors that were included were those listed in Table 2 below, which also provides some statistics. The leftmost number is the number of degrees of freedom the variable had, followed by its significance (or p value), and its R value, which is a measure of the partial correlation between the particular independent variable and the dependent variable.

Table 2: Result of a step-forward logistic regression analysis of independent factors on liaison

Variable	df	Sig (p)	R
Age	2	.0017	.0499
Liaison Frequency	4	.0000	.1190
Liaison Word Class	9	.0000	.1641
Liaison Orthograpy	12	.1263	.0000
Liaison Syllables	5	.0000	.0798
Next Word Class	10	.0000	.1281
Next Word Orthography	4	.0000	.1143
Pause	3	.0276	.0297
Person	3	.0422	.0249
Plural	2	.0002	.0609
Preceding Segment Type	2	.0016	.0501
Punctuation	2	.0000	.0939

We can see from these results that aside from the spelling of the liaison consonant ("Liaison Orthography"), all the factors included are significant with a $p < .05$. Why the liaison consonant's orthography was included despite its lack of significance is clear from the selection criteria. When a factor would significantly increase the power of the model, it was added to it. When LIAS-ORTH was included in the model, the model's reported its predictions to have improved by a full 1%, an improvement significant to a degree of $p < .001$ (not shown here).

Figure 1: Liaison use as a product of age



The only social factor included in this model is the social factor age, which confirms Ashby's (1981) and Booij and De Jong's (1987) finding that age correlates with liaison use. Just as in those studies, older speakers used liaison

more than younger speakers in the current corpus, as shown in Figure 1, above. Interestingly, level of highest education and sex are not found to be significant in this sample (although they do display slight trends).

I then performed a logistic regression analysis for a number of potential interactions between factors. Each trial started with the model described in Section 5 as the base model. Given this model, a step-forward selection procedure was run on the interactions in question. Four of the resulting models included, in addition to the 12 autonomous factors constituting the base model, an interaction between factors. Those interactions that were included, and partial results from the statistical tests are seen in Table 3.

Table 3: Partial results of step-forward logistic regression analysis of interacting and independent factors on liaison in the Liaison Corpus.

Variable	Df	Sig	R
Age * Education	6	.0075	.0618
Age * Liaison-Word-Class	16	.1821	.0000
Liaison-Word-Freq * Next-Word-Freq	16	.0305	.0000
Liaison-Word-Class * Next-Word-Class	46	.0445	.0000

These additions to the model led to a statistically significant increase in the model's predictive power, from 88.3% accuracy to 90.8% accuracy, as well as in several other metrics: log likelihood and goodness of fit. Let's take a closer look now at one of the interactions that were included in this model. The first, third, and fourth interaction terms, although interesting on their own right, are not addressed further here. The focus will lie on the second term, which was selected for further analysis.

The interplay of age and liaison word grammatical class is included as a significant term. In the corpus, young speakers omit more liaison consonants when the liaison word is a verb than when it is an adverb. By comparison, middle-aged and older speakers exhibit little difference in their treatment of these two classes. This trend might indicate a dropping off of the morphological uses of liaison and an increase in its lexicalized uses, as with adverbs.

To sum up so far, through our corpus analysis, we have ascertained that there exists in the language community an interaction between the age of a speaker and the part of speech of the liaison word on the probability that a liaison consonant will be produced. A perception experiment described in the next section addresses the status of this interaction effect in individual language processing.

A Perception Experiment

The statistical analysis above shows one autonomous social factor on the production of liaison, age, as well as interactions between age and two other factors, education and the part of speech of the liaison word. It is the autonomous effect of speaker age and the interaction that crosses the age of the speaker with the identity of the liaison consonant that we will consider in the rest of this section.

As seen in Table 4 below, the proportion of liaison consonants produced in adverbs to those produced in verbs is greater for young speakers than for middle-aged and older speakers. So if subjects have internalized this tendency, then they should judge speakers to be relatively younger if they produce liaison with adverbs rather than not, relative to verbs produced with liaison versus those without.

Table 4: Liaison valence as a product of grammatical class and speaker age

	Liaison word grammatical class	
Age	Adverb	Verb
Old	52%	51%
Middle	38%	39%
Young	43%	37%

Following Bates et al. (1996), I took one of the independent variables, speaker age, and used it as the dependent variable. Adopting this approach allowed the following questions to be addressed:

1. When making social judgments about speakers, do listeners make use of the speakers' production of liaison consonants?
2. Are these judgments contingent upon the interactions between the production of the liaison consonant and other variables, like the grammatical class of the liaison word?

Stimuli

36 words pairs consisting of a liaison word and a following word were selected from the corpus described above. These stimuli varied along three factors - the age of the speaker, the grammatical class of the liaison word, and whether or not the liaison consonant was produced. The 36 stimuli were distributed among the resulting eight conditions as shown in Table 5, below. Of these, only the stimuli produced by the middle-aged speakers were to be subjected to analysis, as subjects were expected to have a great deal of difficulty in correctly discerning the age of speakers given the conditions (this was also established retroactively through a post hoc analysis). Older and younger speakers were included alongside the Middle-Aged test stimuli to ensure a range of speaker ages in the stimulus set.[1]

[1] The classification of speakers who were at the time of experimentation aged 32 to 51 years as "Middle Aged" may be contested. In the author's experience, it frequently is contested, particularly by commentators aged 32 to 51 years of age. The lines were drawn as they were merely such that one-third of the speakers in the corpus fell into each category, not for more contemptible reasons.

Table 5: Distribution of tokens among three factors – age of the speaker, grammatical class of the liaison word, and whether or not liaison was produced.

	1924-1949		**1950-1969**		1970-pres	
	Adv	Ver	**Adv**	**Ver**	Adv	Ver
+liaison	3	3	**6**	**5**	0	2
-liaison	1	1	**6**	**7**	2	2

36 filler stimuli were matched with the test stimuli for length and frequency. Test stimuli were also matched to each other for the factors found to be significant in the corpus study above.

Analysis

From the campus community of the University of Lausanne, 63 subjects were recruited and were compensated with 10 Swiss Francs each, at the time of experimentation equal to about $6, for their participation in this and one other experiment. They ranged in age from 19 to 38 and all self-reported as native speakers of Swiss French. Their regional origins were spread throughout French-speaking Switzerland, but most were born in the Vaud canton, where Lausanne is located. Subjects were told that they were participating in a perception experiment, and, after detailed instructions on the experiment process, were instructed to respond as quickly as possible with their responses to the stimuli. Before the actual recorded part of the experiment began, they engaged in a training session, made up of twelve stimuli very similar to the test and filler stimuli. This experiment lasted no more than ten minutes.

Subjects performed an age-identification task, in which they heard a spoken stimulus, which was selected from either the test or the control stimuli. Their task was to decide, as quickly as possible, whether the speaker of the stimulus was "young", "middle", or "old".

Results

Responses to this forced-choice age-judgment task can provide a window into unconscious cognitive processing. Under extreme time pressure, subjects make semi-automatic judgments, which cannot be entirely due to conscious processing. On average, subjects responded 835 msec after the end of the stimulus.

When, as planned, we consider only the middle aged speakers, precisely the interaction we expected emerges (Table 6).

Table 6: Age judgments for middle-aged speakers as a function of liaison valence and grammatical class.

	Liaison		No Liaison	
	Adverb	Verb	Adverb	Verb
Young	45%	37%	36%	26%
Middle	43%	38%	53%	61%
Old	12%	25%	11%	13%

The proportion of 'young' judgments to 'old' judgments in adverbs where liaison is produced is much greater than the proportion of 'young' judgments to 'old' judgments in verbs where liaison is produced. Similarly, adverbs with produced liaison consonants garnered many more 'young'' judgments than did their counterparts with liaison consonants that went unproduced.

The statistical significance tests of these data were conducted in the following manner. 'Young' and 'old' responses were separately compared with 'middle' responses. In both a forward and backward stepwise model building logistic regression procedure. Thus, there were four tests of significance. In three of the four, all but the backward 'young' to 'middle' comparison, the interaction of valence by grammatical class was deemed significant, with $p < 0.05$. In the other case, it was not included in the model. Grammatical class and valence were also included in a subset of the models.

To summarize, the majority of logistic regression models (three of four) include a term representing the statistically significant interaction between liaison valence and liaison word grammatical class. In other words, there seems to be a statistically significant interaction between whether liaison was produced on a word and what the grammatical class of that word was.

Probability in processing and acquisition

The results described above suggest that individual listeners are unconsciously aware of the correlations between liaison on the one hand and several syntactic and social factors on the other. The age perception task demonstrated the relevance of interactions between factors on liaison for language processing. Other work (Bergen, 2001) has shown through a cross-modal matching paradigm that speakers also have knowledge of statistical correlations between the use of liaison on one hand and liaison segment identity, liaison word length, and liaison word frequency on the other.

These results bear interesting implications for the study of language acquisition. Adult language users encode subtle statistical correlations. In the case shown above, they seem to encode knowledge about the co-occurrence probabilities relating a morpho-phonological process with syntactic and social information about the context in which that process can be applied. We may ask, what learning mechanisms must a learner bring to bear on the language-learning task in order to come to possess such knowledge?

It seems obvious that some sort of implicit statistical learning must be responsible for the acquisition of correlational statistical knowledge. This learning cannot be based on explicit observation or instruction, since no speakers of French are consciously aware of the detailed statistics of liaison, age, and part of speech that they demonstrate unconscious knowledge of. Nor can it be based solely on the speaker's own intuition about or observation of his/her own production behavior. Since age is a variable which can (unfortunately, it seems) not be strictly manipulated by a speaker, and since the great majority of

subjects in the experiment described above fell into the "young" class, they must have been relying on observations of others' speech.

There is growing evidence that children are able to encode conditional probabilities, and to make use of them in various tasks such as word segmentation (Saffran, Newport, & Aslin 1996). From a functional perspective, we can ask what good learning correlations like the ones documented above would serve. Are they simply a burden of the demonstrated human capacity to encode statistical knowledge about the environment, or do they also aid the language user in acts of communication?

Knowing conditional probabilities for the behavior of linguistic units given social facts about the speaker may benefit the hearer in processing the language that speaker produces. If older speakers are more likely to produce liaison in particular contexts, then having access to this knowledge makes predicting and identifying properties of the linguistic signal just a little bit easier. The benefit may only be measurable in milliseconds, but it may be measurable.

Conclusion

That language users display knowledge of subtle, probabilistic correlations between phonology and other domains of knowledge runs counter to various assertions that have been made on the modularity of language and of linguistic sub-modules. Linguistic and psychological theories often view the human language capacity as modular and deterministic. In the study described here, we see that at least some language knowledge is integrated and probabilistic.

Learning a language involves much more than learning the structures of a language. It has been shown here that knowing the statistics of how linguistic and extralinguistic features correlate is part of what must be learned.

References

Ashby, W. (1981). French Liaison as a sociolinguistic phonemenon. In W. Cressey. & J. Napoli (Eds.) *Linguistic Symposium on Romance Languages 9*. Washington, DC : Georgetown UP: 46-57.

Bates, E., Devescovi, A., Hernandez, A., and Pizzamiglio, L. (1996). Gender Priming in Italian. *Perception and Psychophysics*, *58*(7): 992-1004.

Bergen, B. (2001). *Of sounds, mind, and body: Neural explanations for unruly phonology*. Doctoral dissertation, Department of Linguistics, University of California, Berkeley.

Biber, D. Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Bod, R., Hay, J. & Jannedy, S. (Eds.) (2002). *Probabilistic linguistics*. MIT Press.

Booij, G. and de Jong, D. (1987). The domain of liaison: theories and data. *Linguistics 25*: 1005-1025.

Chollet, C., Chochard, J.-L., Constantinescu, A., Jaboulet, C. & Langlais, P. (1996). *Swiss french polyphone and polyvar: Telephone speech database to model inter- and intraspeaker variability.* (Technical Report RR-96-01) IDIAP, Martigny.

Encreve, P. (1988). *La liaison avec et sans enchainement: phonologie tridimensionnelle et usages du francais.* Paris: Editions du Seuil.

Jurafsky, D. (2002). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy, (Eds)., *Probabilistic linguistics*. MIT Press.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Pierrehumbert, J. (2001). Stochastic phonology. *GLOT 5, 6,* 1-13.

Rand, D. & Sankoff, D. (Ms.). *GoldVarb Version 2: A Variable Rule Application for the Macintosh.*

Saffran. J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, *44*, 493-515.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996) Word segmentation: The role of distributional cues. *Journal of Memory and Language 35*.606–621.

Sankoff, D. (1987. Variable Rules. In Ammon, Dittmar & Mattheier (Eds.) *Sociolinguistics: An international handbook of the science of language and society, Vol. I*, 984-997.

Tranel, B. (1981). *Concreteness in generative phonology: evidence from French*. Berkeley: University of California Press.