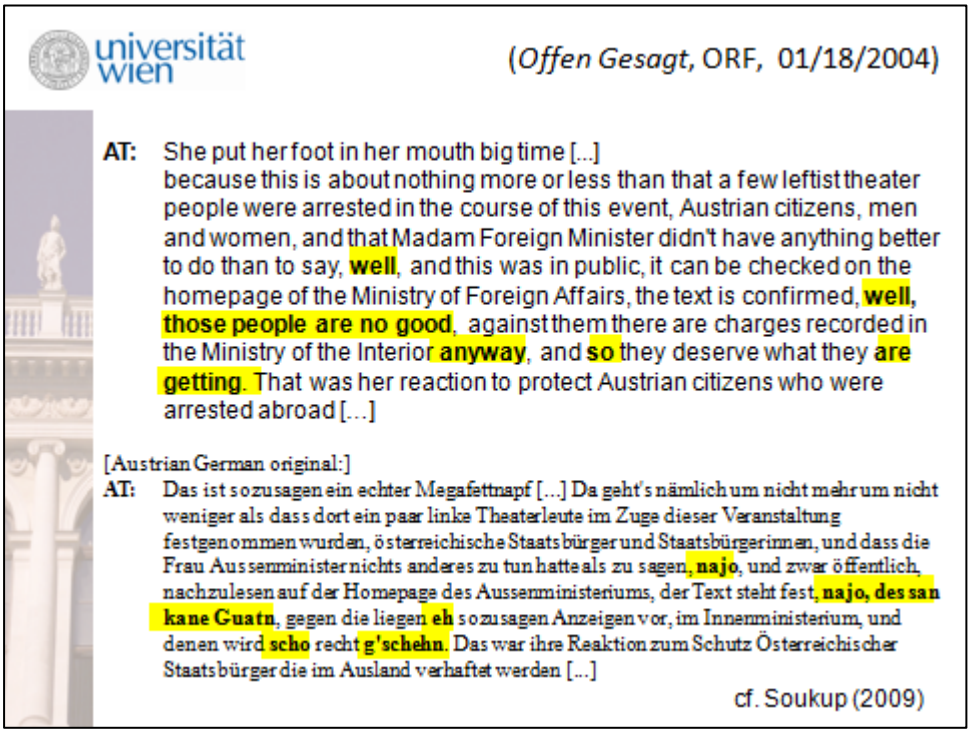
NWAV 40, Georgetown University, October 27-30, 2011

Methodological issues in speech perception elicitation

Barbara Soukup, University of Vienna (barbara.soukup@univie.ac.at)

In my paper today, I would like to discuss three variations of a type of folk linguistic field method that I have found very useful in investigating non-linguists' perceptions of style-shifting in naturally occurring conversation. I have come to call this 'speech perception elicitation'. The first variation is the original, which I carried out six years ago, and then I will present two later retakes with some of the same informants, but using different response schemes, to test which might be the best one.

To begin, I'll show you some data to point out the problem this methodology addresses. Speakers of Austrian German can be found to style-shift routinely between standard Austrian German and Bavarian-Austrian dialect to create rhetorical effects. (One of my colleagues in Vienna actually claims that Austrians do rhetorical style-shifting more than any other group of German-speakers.) You can find this for example in the following data passage taken from an Austrian TV political discussion show called *Offen gesagt*. Here, the speaker is in the process of denigrating his political opponent, the Austrian Foreign Minister. He is recounting an incident where an Austrian alternative theater group was arrested in the course of demonstrations in Italy. And he is claiming that the Austrian Foreign Minister made a big mistake because she did not immediately intervene with Italian authorities to get the theater group out of jail.



Notice that AT is shifting into Bavarian-Austrian dialect in the passage where he is allegedly quoting the Minster – the dialect features are the ones that I've highlighted. At least, that is my linguist's analysis of AT's utterance, based on the relevant literature on Austrian German. I furthermore claim that the style-shift *is* rhetorical, because it *embodies* AT's contempt for the Minister. In Austria, dialect use is commonly associated with stereotypes of low education, low intelligence, and coarseness, as language attitude research has shown. AT's use of dialect in the words he puts into the minister's mouth bring these stereotypes to bear on her as the alleged speaker, so that she is being presented in a very bad light indeed.

So here's the problem now though. In order for AT's strategic language use (or Speaker Design, as Schilling-Estes 2002 calls it) to work, his audience must perceive that a style-shift actually has occurred. *His audience* – here, presumably, the general Austrian TV-watching public - must realize that he is now speaking dialect, not just me as a linguist. My own claims may be based on detailed descriptions of production data, but the fact is that Austrian standard and dialect are very close linguistically, or they actually overlap, and at times it is very difficult to make the call of where shifting occurred. Also, as we know by now, production, or the way people speak, does not automatically equal perception, or what people hear as meaningful linguistic differences.

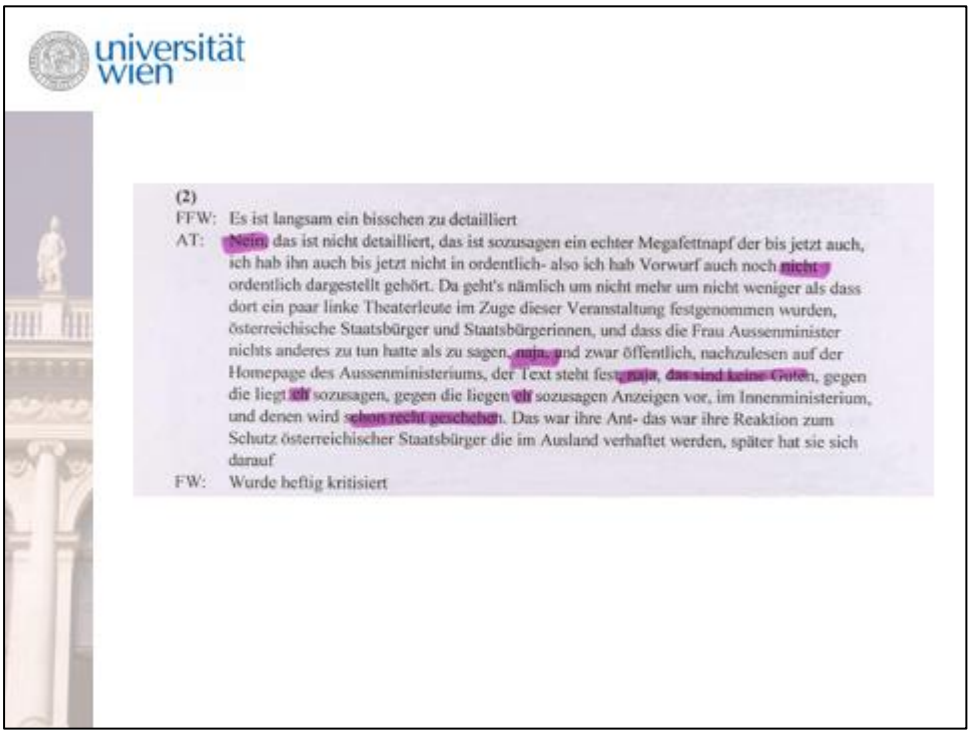
So what is needed here is some kind of methodology that elicits people's perceptions of style-shifts. What I also wanted from this methodology is that it uses *naturally occurring* speech, and not some kind of manipulated samples, because I wanted to get as close as possible to real-life situations such as the experience of watching and interpreting language use on a TV show.

Another consideration was that I wanted to elicit a fairly complete *collection* of features that *each and all*, to a majority of Austrian listeners, evoke the style of 'dialect'. So I did not just want to test individual features one-by-one, as socioperceptual research typically does, in respectively manipulated samples. Furthermore, in much socioperceptual research, perceptual differentiations are typically inferred *indirectly* from the differences informants make in social evaluations of the samples. I on the other hand wanted a more direct approach, where I asked people to indicate style-shifts to me as they heard them.

So here's what I did. This protocol was first presented by Nicholas Coupland in a 1980 article on Welsh English.

For my part, I took 12 passages from episodes of the TV show *Offen gesagt*, like the one I showed you earlier, and played them to native-speakers of Austrian German, asking them to indicate any sequence where, according to their own perception, shifts into Austrian *Dialekt* occurred, as opposed to standard or *Hochsprache*, which is the *expected* variety on the TV show. The selection and designation of these two variety categories is based on ethnographic evidence, that these are indeed the two main meaningful socioperceptual categories in Austrian German.

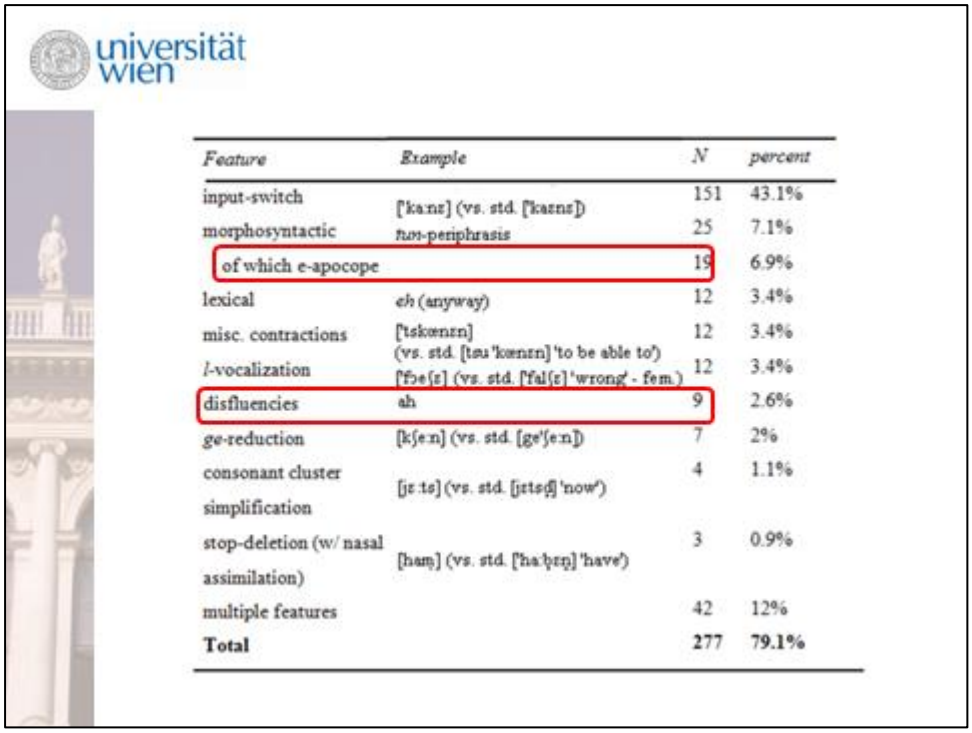
The twelve passages I used were between 35 and 100 seconds long. Each speech sample contained at least a couple of features of Bavarian-Austrian dialect according to the literature.
To record the responses during the task, I gave my informants transcripts of all the excerpts written out entirely in standard German so as not to anticipate any judgments, and they used colored markers to underline any text passage where they heard dialect being used.
A total of 42 informants completed the task. These were recruited from my own family and friends; they ranged in age from 20 to 70. They are all from the Bavarian-Austrian dialect region. Most of them have a middle-class background; about half hold a university degree. With this background, my sample actually corresponds pretty well to the target audience for the TV show *Offen Gesagt*, according to its marketing profile.
The protocol was applied in a total of 19 sessions, with 1-5 participants; each speech sample was played twice in immediate succession.
Now, here's just an example of what the marked-up transcripts ended up looking like:


universität wien

(2)
FFW: Es ist langsam ein bisschen zu detailliert
AT: Nein das ist nicht detailliert, das ist sozusagen ein echter Megafettnapf der bis jetzt auch, ich hab ihn auch bis jetzt nicht in ordentlich- also ich hab Vorwurf auch noch nicht ordentlich dargestellt gehört. Da geht's nämlich um nicht mehr um nicht weniger als dass dort ein paar linke Theaterleute im Zuge dieser Veranstaltung festgenommen wurden, österreichische Staatsbürger und Staatsbürgerinnen, und dass die Frau Aussenminister nichts anderes zu tun hatte als zu sagen, naja, und zwar öffentlich, nachzulesen auf der Homepage des Aussenministeriums, der Text steht fest, naja, das sind keine Guten, gegen die liegt ehr sozusagen, gegen die liegen ehr sozusagen Anzeigen vor, im Innenministerium, und denen wird schon recht geschehen. Das war ihre Ant- das war ihre Reaktion zum Schutz österreichischer Staatsbürger die im Ausland verhaftet werden, später hat sie sich darauf
FW: Wurde heftig kritisiert

For my analysis, then, I first tabulated how many times each word in the transcript had been underlined by the informants.
I actually set a cut-off at words underlined by at least a quarter of informants, because I wanted to get a *consensual* list of features that the general public would perceive as dialectal – a sort of common denominator. Out of a total of over 2,000 words in the transcript, exactly 350 were underlined by at least a quarter of my informants. As a next step, I matched these words with close IPA transcriptions of their actual realizations in the samples, in order to find the reasons for which the words might have been underlined. For this, I did rely on descriptions of salient

features of Austrian dialect as established in respective literature. Here's a table summarizing the types of features I was able to detect in those words that had been underlined by at least a quarter (or 11) of my informants.


universität wien

Feature	*Example*	*N*	*percent*
input-switch	['ka:nɪ] (vs. std. ['kaɪnɪ])	151	43.1%
morphosyntactic	*tun*-periphrasis	25	7.1%
of which e-apocope		19	6.9%
lexical	*eh* (anyway)	12	3.4%
misc. contractions	['tskœnɪn] (vs. std. [tsu 'kœnɪn] 'to be able to')	12	3.4%
l-vocalization	['fɒɐ̯ɛ] (vs. std. ['fal(ɛ)] 'wrong' - fem.)	12	3.4%
disfluencies	eh	9	2.6%
ge-reduction	[kʃe:n] (vs. std. [ge'ʃe:n])	7	2%
consonant cluster simplification	[ɪs:ts] (vs. std. [ɪʃtsɐ̯ɪ] 'now')	4	1.1%
stop-deletion (w/ nasal assimilation)	[ham] (vs. std. ['ha:bɪn] 'have')	3	0.9%
multiple features		42	12%
Total		**277**	**79.1%**

By far the biggest proportion of words identified as dialectal contain 'input-switches', meaning, forms in which the difference between standard and dialect is due to a diverging historical development; n=151), followed by words with multiple features (n=42), morphosyntactic features (n=25), lexical features, contractions, and *l*-vocalizations (n=12 for each). A crude calculation shows that each occurrence of a word showing one of these features in the data set was on average underlined by more than half of the informants (55%).

Now what I believe this protocol has conveniently allowed me to establish, on an empirical basis, is a list of features that Austrian listeners will most likely perceive as constituting shifts into dialect. Virtually none of the features on this list occur in those parts of the transcripts that were *not* underlined by *any* informant, so this seems pretty solid.

Yet one can of course argue that by virtue of providing my informants with written standard transcripts, I was putting too much focus on written language as a benchmark. (And in Austrian German this is always an issue at any rate, because there is no clear spoken standard as an accepted norm of reference.)

So we find, for example, that 19 of the highly underlined words in the transcript show the feature of e-a'pocope, or deletion of the verbal inflection of 1st or 3rd person singular present tense: e.g. *(ich) mein* vs. written standard *(ich) mein**e*** – '(I) mean'. Such deletion of the suffix is categorical in the dialect, but it is also widely attested in upper class/ educated/ formal speech– so, in

language use by speakers and in situations that might traditionally be considered spoken standard. So this is one particular point where it is not clear whether the *written* response scheme might have overly influenced responses to lean towards the written standard, and perceive any features of oral speech production automatically as dialectal.

Similarly, a very interesting thing I found in my data is that my informants had also marked speech disfluencies, hesitations, and false starts as dialectal. Probably no linguist would *classify* these as dialectal, but they would rather be considered 'normal' features of spoken vs. written language. So here, too, the written response scheme might have played a decisive role.

So in order to find out whether the response scheme hyperdefined the referent as being the *written* standard, and not any form of realistic spoken standard, I tried another tack. The original test was conducted back in 2005. Five years later, in 2010, I was able to recontact six of my original informants to do a retest, this time without the transcripts.

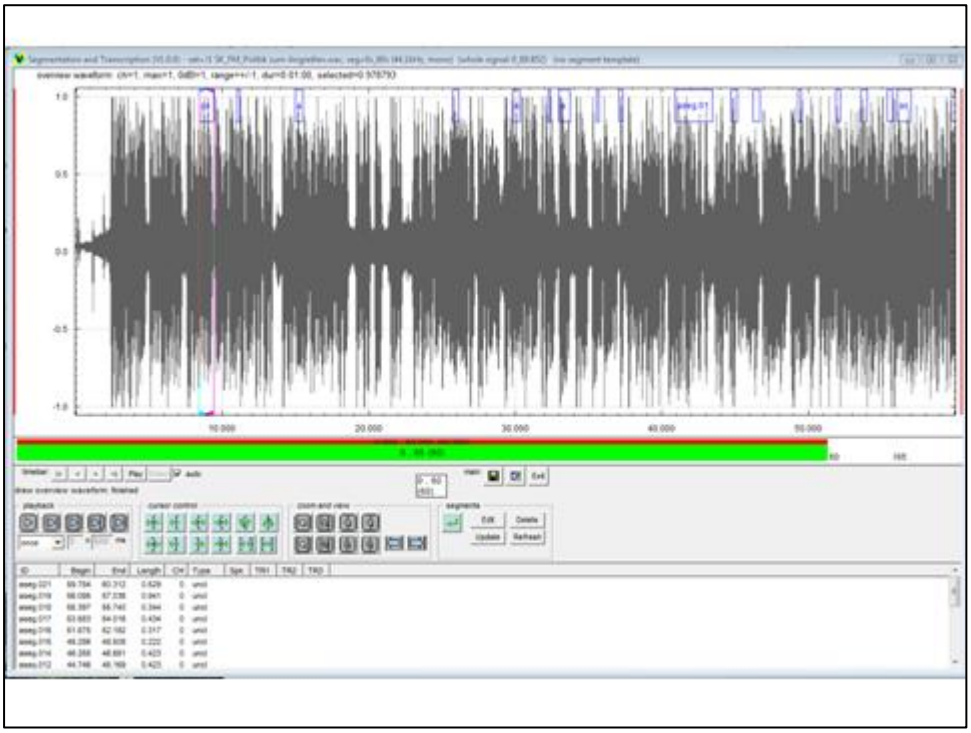
Instead, I asked my informants to listen to the same recordings again (they said they didn't remember anything), and to *themselves* transcribe by hand what they heard, in eye dialect, where applicable. The idea was that this would have them focus more on the audio mode, and get further away from a written baseline.

The procedure was as follows. I had my informants listen to the audio on my computer, and transcribe in handwriting while I was handling the media player and stopping and backing up as much and as often as they wished. So this was a pretty drawn-out process, and we can argue about how natural it still actually is in terms of capturing speech processing. Once we had thus arrived at the end of one sample, we did a last run-through of the whole to check it one more time. Then I asked my informants again to underline in their own transcript any words they perceived as *Dialekt*, just like in the first experiment, and at the same time I made sure I was able to decipher their handwriting and use of eye-dialect, because I later had to type up and code the transcriptions for my analysis. So we did all this for one audio sample, then moved on to the next.

You can probably already imagine some of the complexities of this approach. First of all, as we all know, transcribing audio is a very tedious task, and one that takes some getting used to. My informants were all non-linguists, but they actually mastered the process surprisingly well. But in this set-up, the experiment also took much longer – up to 90 minutes, as opposed to the 45 min from before, even though I had cut the samples a bit and taken out three with very fast speech rate. The speed of the natural speech samples already was an issue in the first version, but even more so in the second, where the informants had no indication what was coming at them content-wise – none of them remembered. Also, I could only do the second experiment with one informant at a time, while I'd done the first in groups.

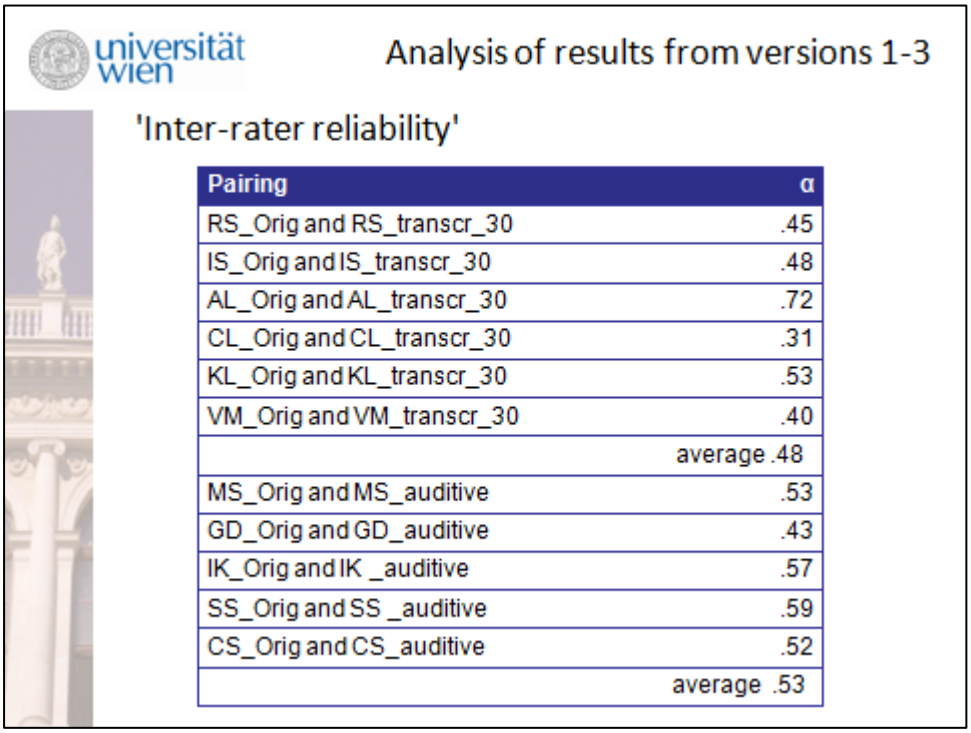
So all in all I have to say that the first response scheme, with the written transcript provided, was decidedly more informant- and researcher-friendly. I'll get to a comparison of results in a minute, after presenting my second retake version of the speech perception elicitation test, carried out earlier this year.

My first plan for this was to have another set of my original informants mark up an oscillogram of the audio samples they were hearing. This was suggested to me by Christoph Draxler of the University of Munich. I tried this with one informant who was very computer-savvy, using the annotation features of the software STX. Here's what this ended up looking like.



As it turned out, though, this was a very unwieldy and ineffective method. It took my informant 45 minutes to just get through one segment, and later he reported that by virtue of the micro-analysis and close listening needed to carry out the task, he had trouble keeping a sense of what the speakers were actually saying. This was getting fairly far away from any type of ad hoc natural speech processing. So what I ended up doing with five more informants was to have them listen to the samples and just indicate to me orally any passages where they perceived the use of dialect. They first listened to each passage once in its entirety, and then proceeded through it again in a stop-and-go fashion. I sat next to them and wrote down what they were pointing out as dialectal. We then finalized the list of words that they had identified and moved on. This procedure was as fast as the first one with the written transcripts (though I could of course also only proceed one informant at a time). As informants became familiar with the task, they started to move quite speedily through; and they did not report any particular problems, which is good. So the question now is how did the results from these three different protocols match up.

First off, the inventory of features I found across the different response schemes remained pretty much the same. Now, to see if words containing these features were also underlined consistently, I calculated inter-rater reliability for each of the informants between either the original and the transcribing task, or the original and the oral response, using Krippendorff's alpha.



Pairing	α
RS_Orig and RS_transcr_30	.45
IS_Orig and IS_transcr_30	.48
AL_Orig and AL_transcr_30	.72
CL_Orig and CL_transcr_30	.31
KL_Orig and KL_transcr_30	.53
VM_Orig and VM_transcr_30	.40
	average .48
MS_Orig and MS_auditive	.53
GD_Orig and GD_auditive	.43
IK_Orig and IK _auditive	.57
SS_Orig and SS _auditive	.59
CS_Orig and CS_auditive	.52
	average .53

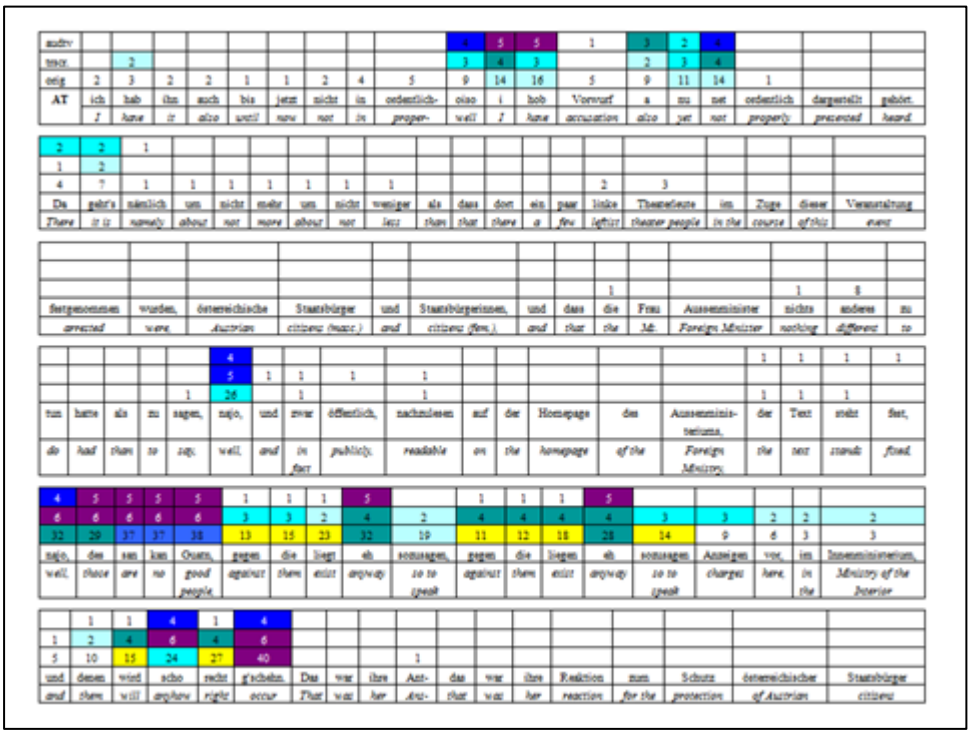
Reliability was 'fair', around 50% on average, and this average did not differ significantly across the two alternative response schemes. As for the average amount of words that the informants underlined between the original and one of the retakes, there was no difference within the group that did the transcribing retake, but those informants that did the oral response scheme actually pointed out more features overall. Mind you, I am using the aggregated data here – there are certainly idiosyncratic differences between informants in terms of amounts of marking, but my main interest, as I said before, lies in finding the common consensus.

Also, in terms of types of features, both groups pointed out more input-switches in either retake than in the first round, and the oral response group also pointed out more lexical features, l-vocalizations, and words with multiple features. I would attribute all this to the fact that both response schemes for the retakes involved a closer analysis and more time spent listening and deciding than the original written response task. I tentatively suggest that this might mean that the original got a little closer to natural speech processing, though the actual process must be infinitely faster than we can ever hope to capture and make explicit.

There were no other significant differences in terms of the average amount of words underlined by feature category. This is important to note, because I said earlier that it was unclear to me after the original study whether the fact that informants had underlined speech disfluencies and

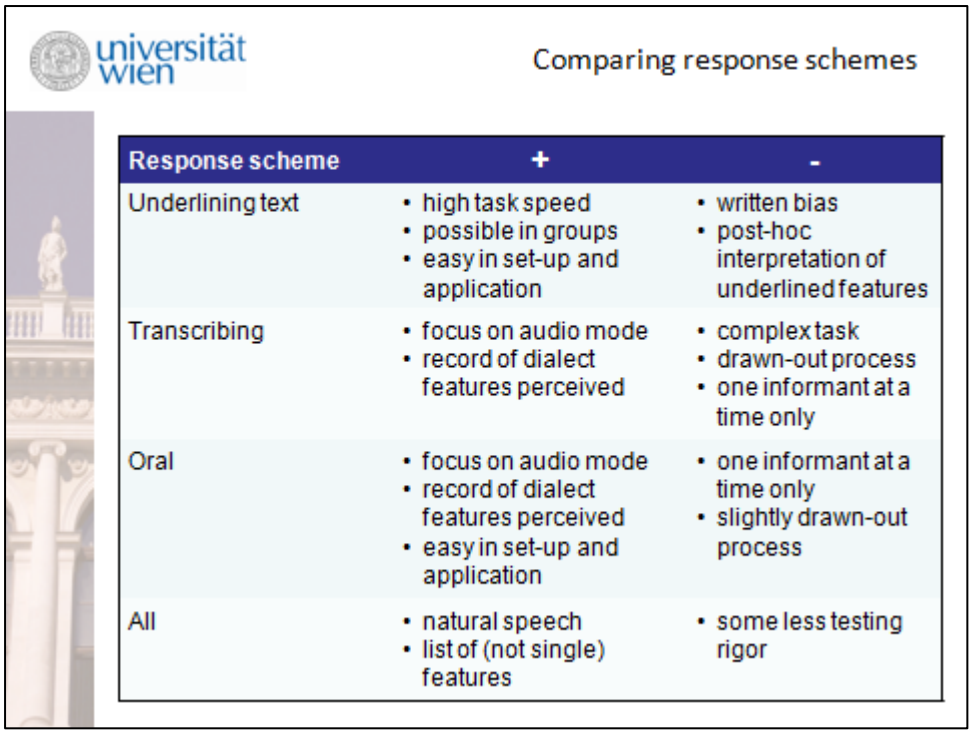
e-apocope as dialectal was a function of the response scheme or no. But as it turns out, there is now no clear pattern under which either of these had gotten marked more under the written scheme. I believe this confirms what other people have claimed before me, that in Austrian German, there is very generally an orientation towards the written standard, for lack of a spoken standard pronunciation norm, and this comes out either way you test it.

The low cell counts I have with my limited number of retest informants don't allow me to do much further statistical testing. But what I can show you as my main finding is the following:



If you match up all those cases across the three test conditions in which at least a quarter of the respective informant group underlined a word as dialectal, the picture is very similar. You can see here the excerpt that I showed you at the beginning, about the Foreign Minister. The darker the shading gets, the more informants underlined a word as *Dialekt*. (Top is oral response, second is transcribing, third is written response scheme). You can see here that all three response schemes, on a global level, detected pretty much the same 'hot-spots' of standard-dialect shifting that a discourse analysis should probably focus on. So overall, the results seem quite robust, if what one is looking for is indeed a common denominator of perceptions of style-shifting.

Finally, here's a quick look at the response schemes themselves and their affordances and constraints.



The underlining scheme was fastest in task speed. You are putting pressure on informants to give a first, intuitive response, and that is probably an affordance. On the downside, you'd really have to test for a written language bias again if you apply this to a different context. Also, my interpretation of the underlined features was admittedly carried out post-hoc, on the basis of the literature and my own perception, because all I have as a record is an underlined word, and no reason why. Chances are that informants went beyond those established features in some cases and had some of their own reasons for underlining words.

Second, the *transcribing scheme* has on its plus side a stronger focus on the audio mode, and that informants are themselves creating a record of the dialect features they hear. On the downside, it is a very complex set-up that is drawn out and thus perhaps not tapping into people's first response, and unlike the written scheme, one can only proceed one informant at a time.

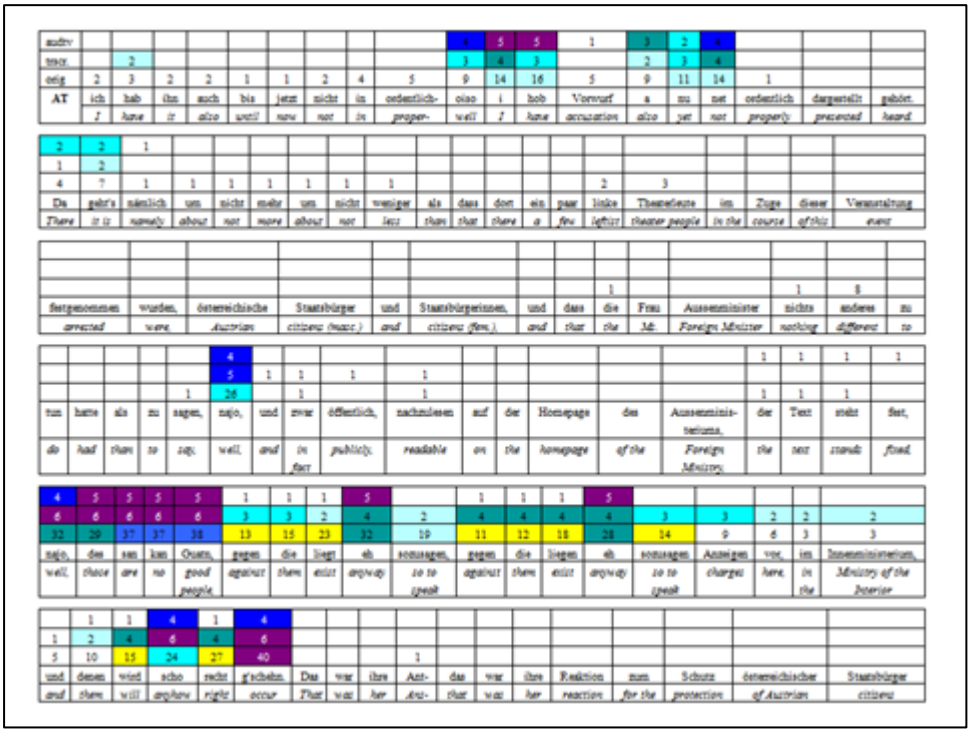
The oral response scheme put the focus even more on listening only, because the informants did not have to write anything down – I was doing that. You also get a record of dialect features, and it is easy to set up. On the other hand, it also only works one informant at a time, and it is arguably more drawn-out and microanalytic than the written scheme.

All in all, then, while the original written response scheme might get closest to fast, online speech processing in natural conversation of the three, there is actually much to be said for the oral response scheme, because it is also easy to use, almost as instant, and in addition yields valuable information regarding *why* informants perceive certain words as dialectal.

One affordance that all these types of speech perception elicitation tests share, is their use of natural speech samples, and the fact that one can elicit a consensual collection of linguistic features that are perceived as constituting a certain stylistic category, and not just test single

features that are artificially manipulated. I do believe that this makes up for the certain lack of testing rigor involved. At the very least, traditional socioperceptual research and speech perception elicitation such as I have described it should be able to complement each other well.

And to conclude with my previous slide,



any and all of these response schemes I have applied were able to pick out the same hot-spots of style-shifting. Picking out these hot-spots perceptually is very necessary, if we want to make claims about the communicative effects of strategic language use, or Speaker Design. So I would argue that speech perception elicitation tests are an important addition to our empirical toolkit, that reminds us, and allows us, to pay due attention to the listener as an equal partner in the activity of sociolinguistic meaning-making via style-shifting.

THANK YOU!