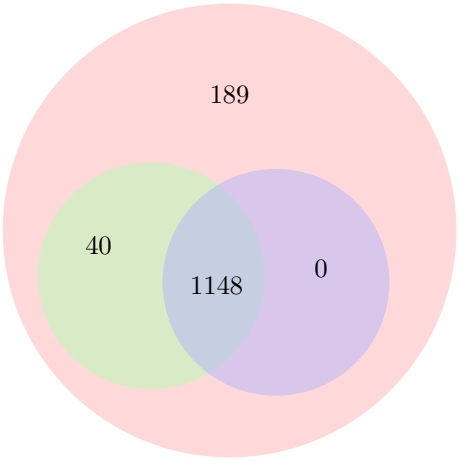
Typology I: Solution to Homework for Lecture 11

1. a) A *convenience sample* includes languages for which there is a grammatical description of the feature under investigation (regardless of the exact feature value). Hence, the biggest convenience sample contains all languages in bold face.
   b) Given the answer in a), we know that this sample can only contain languages in bold face. Additionally, *phylogenetically balanced* means that we are not over-representing any family by sampling more languages of that family than of any other. Since the Eskimo-Aleut family is only represented by one language (Yupik), we can only draw one language from each of the other families. Hence, one possible sample would be French, Maori, Hebrew, Yupik.
   c) This is strictly speaking not possible. A *variety sample* should contain languages representing all the possible values of a feature, i.e. f1, f2, f3, f4 and f5 in our case. Note that in order to represent f3 and f4 we have to draw two languages of the Austronesian family (Maori and Rarotongan). To phylogenetically balance the sample, we would have to sample two languages from all the other families. However, this is not possible for the Eskimo-Aleut family, since only Yupik is described. Hence, we either have to accept that the sample is not completely balanced, or include Inuktitut with the value of the feature set to NA.

2. * a) There are 31 languages in total. 17 languages have no tone (1st category), 14 languages have tone (either 2nd or 3rd category). Hence, the overall probability for having tone in this sample is $\frac{14}{31} = 0.45$ or 45%.
   b) The smallest family represented is Khoisan with 4 languages. Hence, 4 languages is the maximum we can draw from the other families to not bias the sample. We arrive at 4× *3 Complex tone system* (Oto-Manguean) + 4× 1 No tones (Uto-Aztecan) + 3× *2 Simple tone system* (Khoisan) + 1× 3 *Complex tone system* (Khoisan) + 4× 1 No tones (Austronesian). This means there are 16 languages in our sample, of which 8 do not have tone and 8 have tone. The probability is $\frac{8}{16} = 0.5$ or 50%.

3. $Y \cap Z$: All languages which have a dominant word order *and* have SO order, i.e. **1148**.
   Set Y: All languages which have a dominant word order *but do not* have SO order, i.e. **40**.
   Set Z: All languages which have SO order but *do not* have a dominant word order. This is logically impossible, i.e. **0**.
   Set of all languages: The remaining languages. Languages that do not have a dominant word order, i.e. **189**



Further notes: It is important to double-check that all the numbers add up to the overall number of languages, i.e. *1148+40+189=1377*. In terms of its architecture the Venn diagram used in the lecture is representing a scenario where there are some languages in Set Z only. Strictly speaking, the Venn diagram is wrong for these specific numbers from WALS 81A, since Set Z would have to be a proper subset of Set Y, not just an overlapping set.

4. Alice in Wonderland

   - word tokens with two morphemes: *beginn-ing, tir-ed, sitt-ing, hav-ing, (no-thing),*[1] *peep-ed, (in-to), read-ing*
     synthesis index = morphemes/word = 42/34 = **0.98** [2]

   - inflected word tokens: *was, beginn-ing, tir-ed, sitt-ing, hav-ing, had, peep-ed, was, read-ing*
     inflection index = inflections/word = 9/34 = **0.26**

   - derived word tokens: -
     derivation index = derivations/word = 0/34 = **0**

   UDHR, Article 2

   - word tokens with two morphemes: *Every-one, entitl-ed, right-s, freedom-s, Declarat-ion, with-out, distinct-ion, (langu-age), relig-ion, polit(-ic)-al, (opin-ion), nation-al, soci-al*
     synthesis index = morphemes/word = 48/34 = **1.41**

[1] word tokens in parenthesis might be counted as two morphemes or just one, I would accept both. The indexes are calculated by counting them as two (or more) morphemes.
[2] all numbers are rounded to second decimal place

- inflected word tokens: *is, entitl-ed, right-s, freedom-s*
  inflection index = inflections/word = 4/34 = **0.12**

- derived word tokens: *Declarat-ion, distinct-ion, relig-ion, polit(-ic)-al, (opin-ion), nation-al, soci-al*
  derivation index = derivations/word = 8/34 = **0.24**

Interpretation: Both the synthesis and derivational indexes are higher for the UDHR than the Alice in Wonderland passage, while the inflectional index is lower. Note that the number of tokens is the same in both passages, so differences in text size do not account for the deviations. Rather, the deviations are due to extensive usage of derived vocabulary in the UDHR to express complex legal concepts, while the story of Alice is a narration set in the past, which requires verbal inflection to distinguish tense and aspect, e.g. *was beginning*, *had peeped*. Hence, the differences are due to register and style, i.e. formal "legalese" on one hand, and narrative prose on the other hand.

Further notes: The cases in parenthesis such as *(no-thing)*, *(langu-age)* or *polit(-ic)-al* are tricky. Is *nothing* perceived by speakers as a single morpheme bearing a "single" meaning, or as a semantic compound of "no" and "thing". Likewise, if words were already borrowed with derivational morphology into English, such as *politics* from Greek *polis* and *polit-ikos*, then the question is whether this derivational morphplogy is still productive in English usage.