[Taiwan Language Survey](https://twlangsurvey.github.io/" \t "_blank) is a small project I worked on during May to June in 2018. The idea was to create a survey that **continuously** collects data and a web page that visualizes the collected data. The web page is updated weekly using [Travis-CI](https://travis-ci.org/twLangSurvey/twLangSurvey.github.io).

The main purpose of this survey is to raise public awareness of **language loss** in Taiwan. Hence, the survey is designed to collect data that can provide valuable information about language loss, for example, some questions were asked to gain insight about the change of linguistic competence acoss generations in a family (i.e. across the subject’s parents, the subject, and the subject’s children). In addition to changes within family, information about the age of the subjects is also colleceted, meaning that we can see how linguistic competence changes among diffrent age groups of subjects in a community, i.e. is a language becoming more dominant or entering the dying process?

Visualization is a powerful tool to capture how linguistic competences of different langauges are changing. But creating visualizations necessitates creativity – how can language loss be visualized? Below, I illustrate one of the methods I created for visualizing language loss – a visaulization inspired by the [age-sex pyramid](https://en.wikipedia.org/wiki/Population_pyramid).

**Age-Sex Pyramid of Language**

The age-sex pyramid is used to visualize the population structure of a community. The vertical axis indicates the age and each horizontal bar represents an age group. The horizontal axis indicates the population size of male or female of a particular age group.

The age-sex pyramid is a great tool to visualize the population structure since the ‘shape’ of the pyramid gives readers a lot information. For example, an ‘expansive pyramid’ has longer bars at the bottom of the pyramid, which indicates the population is young and growing. A ‘stationary pyramid’ looks like a rectangular bar, indicating the population sizes of different age groups are about the same. A ‘constructive’ pyramid indicates a shrinking population, which is narrowed at the bottom.

Similarly, a **modified** version of the age-sex pyramid, which I’ll call the ‘**age-sex pyramid of language**’, can be used to visualize the population structure of a language and predicts the language’s **vitality**. Instead of visualizing population size, the age-sex pyramid of language visualizes the **average fluency of a language** on the horizontal axis.

An age-sex pyramid of Taiwanese. The red bars on the left indicates females of different age groups and the blue bars on the right indicates males. The average fluency (values on the horizontal axis) is calculated from a six-point scale (0-5) on Taiwanese fluency.

Figure 1: An age-sex pyramid of Taiwanese. The red bars on the left indicates females of different age groups and the blue bars on the right indicates males. The average fluency (values on the horizontal axis) is calculated from a six-point scale (0-5) on Taiwanese fluency.

As shown in Figure [1](https://liao961120.github.io/2019/02/17/visualize-language-loss.html#fig:tw-pyramid), the shape of the age-sex pyramid of [Taiwanese](https://en.wikipedia.org/wiki/Taiwanese_Hokkien) in Taiwan[1](https://liao961120.github.io/2019/02/17/visualize-language-loss.html#fn1) is an ‘**inverted triangle**’, which is almost never seen in the conventional population pyramid. However, this inverted triangular shape is expected to appear quite often, since it indicates an ongoing language loss in a community.

| **Vitality of Language** | **Shape of Pyramid** |
| --- | --- |
| Shrinking and Dying | Inverted Triangle |
| Growing | Triangle |
| Stable | Retangualr Bar |

**Drawing Age-Sex Pyramid with ggplot2**

As complex as it might seem, an age-sex pyramid created with ggplot2 is actually a (modified) bar chart. I learned this on [stackoverflow](https://stackoverflow.com/a/36804394), and the trick is

1. Use ifelse to flip the value (here, population size) according to the gender of the age group
2. Use geom\_bar(stat = "identity") to let the heights of the bars represent values in the data frame, i.e. the value given to y
3. Use coord\_flip() to make the bars horizontal

set.seed(1)

df0 <- tibble::tibble(

Age = rep(c('10-19', '20-29', '30-39'), 2),

Gender = rep(c('Female', 'Male'), each = 3),

PopSize = sample(0:100, size = 6, replace = T)

)

df0

| **Age** | **Gender** | **PopSize** |
| --- | --- | --- |
| 10-19 | Female | 26 |
| 20-29 | Female | 37 |
| 30-39 | Female | 57 |
| 10-19 | Male | 91 |
| 20-29 | Male | 20 |
| 30-39 | Male | 90 |

library(ggplot2)

pl <- ggplot(df0, aes(x = Age,

y = ifelse(Gender == 'Male', PopSize, -PopSize),

fill = Gender)) +

geom\_bar(stat = 'identity') +

coord\_flip()

pl

To center the plot (i.e. to make the point where population size is zero at the center of the plot), we have to scale the axis of ‘population size’ (y) with scale\_y\_continuous:

pl + scale\_y\_continuous(limits = c(-100, 100),

breaks = seq(-100, 100, 25),

labels = abs) +

labs(y = 'Population Size')

where abs in labels is the function abs(). By default, ggplot2 pass the value given in breaks to the the function specified in labels.

**Visualizing Language Loss**

To create an age-sex pyramid of language, the data structure needed is exactly the same as the one above, except the variable, PopSize, is replaced by ‘average fluency’ of a language. But since most people in Taiwan can speak more than one language (e.g. Mandarin-Taiwanese, Mandarin-Taiwanese-Hakka, Mandarin-English, etc.), the real data from the survey is a bit more complex Basically, the data structure needed to draw an age-sex pyramid of language looks like:

| **Gender** | **Ethnicity** | **Age Group** | **Avg. Fluency** |
| --- | --- | --- | --- |
| female | Mandarin | 20-24 | 4.53 |
| male | Taiwanese | 20-24 | 2.78 |
| … | … | … | … |
| female | Hakka | 25-29 | 2.23 |
| male | Taiwanese | 35-39 | 3.57 |

**Preparation of Data**

The raw data of [Taiwan Language Survey](https://twlangsurvey.github.io/) can be retrieved [here](https://docs.google.com/spreadsheets/d/1NOJ9O_zBR6s-9e1fFGVcRaBjn5qQuVB9ZeSebhBctoM/edit?usp=sharing). The survey and raw data is in traditional Chinese. I’ll skip the step of cleaning raw data (e.g., turn variable names to English) and used the cleaned data survey.rds instead.

temp <- tempfile()

download.file('<https://raw.githubusercontent.com/twLangSurvey/twLangSurvey.github.io/master/data/survey.rds>', destfile = temp)

data <- readr::read\_rds(temp)

head(data, 3)

| **date** | **curr\_resid** | **curr\_resid\_since** | **settle\_5yy** | **home\_town** | **gender** | **age** | **kid\_num** | **edu\_level** | **work** | **income** | **work\_hr** | **tribe** | **Mand\_listen** | **Mand\_speak** | **Tw\_listen** | **Tw\_speak** | **Hak\_listen** | **Hak\_speak** | **Ind\_listen** | **Ind\_speak** | **SEA\_listen** | **SEA\_speak** | **Eng\_listen** | **Eng\_speak** | **first\_lang** | **when\_Mand** | **when\_Tw** | **when\_Hak** | **when\_Ind** | **when\_SEA** | **when\_Eng** | **m\_guard\_identity** | **f\_guard\_identity** | **dad\_Mand\_speak** | **dad\_Tw\_speak** | **dad\_Hak\_speak** | **dad\_Eng\_speak** | **dad\_Ind\_speak** | **dad\_SEA\_speak** | **mom\_Mand\_speak** | **mom\_Tw\_speak** | **mom\_Hak\_speak** | **mom\_Eng\_speak** | **mom\_Ind\_speak** | **mom\_SEA\_speak** | **dad\_mom\_Mand\_fq** | **dad\_mom\_Tw\_fq** | **dad\_mom\_Hak\_fq** | **dad\_mom\_Ind\_fq** | **dad\_mom\_SEA\_fq** | **dad\_mom\_Other\_fq** | **me\_dad\_Mand\_fq** | **me\_dad\_Tw\_fq** | **me\_dad\_Hak\_fq** | **me\_dad\_Ind\_fq** | **me\_dad\_SEA\_fq** | **me\_dad\_Other\_fq** | **me\_mom\_Mand\_fq** | **me\_mom\_Tw\_fq** | **me\_mom\_Hak\_fq** | **me\_mom\_Ind\_fq** | **me\_mom\_SEA\_fq**me\_mom\_Other\_fq |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2018-06-12 | 106 | 1996 | 是 | 428 | 女 | 56 | 2 | 大學 | 金融業 | 10萬以上 | 45 – 50 小時 | 不具原住民身份 | 5 | 4 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 華語 | 7歲前 | 7歲前 | 未學會 | 未學會 | 未學會 | 12-15歲 | 父親 | 母親 | 3 | 0 | 0 | 0 | 0 | 0 | 3 | 4 | 0 | 0 | 0 | 0 | 幾乎全用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 幾乎全用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 約一半 | 多數使用 | 幾乎不用 | 幾乎不用 | 幾乎不用幾乎不用 |
| 2018-06-13 | 204 | 2004 | 是 | 204 | 女 | 57 | 1 | 碩士 | 製造業 | 10萬以上 | 50 – 55 小時 | 不具原住民身份 | 5 | 5 | 5 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 4 | 華語 | 7歲前 | 7歲前 | 未學會 | 未學會 | 未學會 | 12-15歲 | 父親 | 母親 | 5 | 5 | 0 | 1 | 0 | 0 | 1 | 5 | 0 | 0 | 0 | 0 | 幾乎不用 | 多數使用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 多數使用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 多數使用 | 幾乎不用 | 幾乎不用 | 幾乎不用幾乎不用 |
| 2018-06-13 | 103 | 2013 | 是 | 100 | 男 | 37 | 2 | 大學 | 金融業 | 55,000 – 60,000 | 25 – 30 小時 | 不具原住民身份 | 5 | 5 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 華語 | 7歲前 | 未學會 | 未學會 | 未學會 | 未學會 | 7-12歲 | 父親 | 母親 | 5 | 5 | 5 | 1 | 0 | 0 | 5 | 3 | 0 | 4 | 0 | 0 | 幾乎全用 | 少數使用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 多數使用 | 少數使用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 幾乎全用 | 幾乎不用 | 幾乎不用 | 幾乎不用 | 幾乎不用幾乎不用 |

The survey data, data, contains 22 variables and 257 observations (subjects). I only need these variables below to create the plot I want:

* gender: The gender of the subject
* age: The age of the subject
* m\_guard\_identity: Male guardian of the subject, should be ‘father’ in most cases
* f\_guard\_identity: Female guardian of the subject, should be ‘mother’ in most cases
* \_speak: The subject’s fluency of a language. is one of ‘Mand’ (Mandarin), ‘Tw’ (Taiwanese), ‘Hak’ ([Hakka](https://en.wikipedia.org/wiki/Hakka_Chinese)), ‘Ind’ (languages of the indigenous peoples, aka [Formosan languages](https://en.wikipedia.org/wiki/Formosan_languages), belong to Austronesian languages), ‘SEA’ (languages from South Easth Asia), and ‘Eng’ (English)
* dad\_\_speak: The subject’s male guardian’s fluency of a language. is same as above.
* mom\_\_speak: The subject’s female guardian’s fluency of a language. is same as above.

library(dplyr)

lang <- c('Mand', 'Tw', 'Hak', 'Ind', 'SEA', 'Eng')

cols <- c('gender', 'age', 'm\_guard\_identity', 'f\_guard\_identity',

paste0(lang, '\_speak'),

paste0('dad\_', lang, '\_speak'),

paste0('mom\_', lang, '\_speak')

)

data <- data[, cols] %>%

filter(gender == '男' | gender == '女')

Since some content of the survey data is in Chinese, the code below is used to translate it to English:

ch2eng <- function(x) {

if (x == '男') return('Male')

if (x == '女') return('Female')

if (x == '母親') return('Mother')

if (x == '父親') return('Father')

if (x == '無') return('None')

if (x == '(外)祖父') return('Grandpa')

if (x == '(外)祖母') return('Grandma')

if (x %in% c('阿姨', '嬸嬸', '舅媽', '姑姑', '伯母')) return('Aunt')

if (x %in% c('叔叔', '伯伯', '舅舅', '姑丈', '姨丈')) return('Uncle')

message('No translation found for `', x, '`', '\n')

return(x)

}

ch2eng\_vec <- function(vec) {

new\_vec <- vector(typeof(vec), length(vec))

for (i in seq\_along(vec)) {

new\_vec[[i]] <- ch2eng(vec[[i]])

}

return(new\_vec)

}

data <- data %>%

mutate(gender = ch2eng\_vec(gender),

m\_guard\_identity = ch2eng\_vec(m\_guard\_identity),

f\_guard\_identity = ch2eng\_vec(f\_guard\_identity))

head(data)

| **gender** | **age** | **m\_guard\_identity** | **f\_guard\_identity** | **Mand\_speak** | **Tw\_speak** | **Hak\_speak** | **Ind\_speak** | **SEA\_speak** | **Eng\_speak** | **dad\_Mand\_speak** | **dad\_Tw\_speak** | **dad\_Hak\_speak** | **dad\_Ind\_speak** | **dad\_SEA\_speak** | **dad\_Eng\_speak** | **mom\_Mand\_speak** | **mom\_Tw\_speak** | **mom\_Hak\_speak** | **mom\_Ind\_speak** | **mom\_SEA\_speak** | **mom\_Eng\_speak** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Female | 56 | Father | Mother | 4 | 2 | 0 | 0 | 0 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 3 | 4 | 0 | 0 | 0 | 0 |
| Female | 57 | Father | Mother | 5 | 5 | 0 | 0 | 0 | 4 | 5 | 5 | 0 | 0 | 0 | 1 | 1 | 5 | 0 | 0 | 0 | 0 |
| Male | 37 | Father | Mother | 5 | 1 | 0 | 0 | 0 | 2 | 5 | 5 | 5 | 0 | 0 | 1 | 5 | 3 | 0 | 0 | 0 | 4 |
| Female | 44 | Father | Mother | 5 | 4 | 0 | 0 | 0 | 4 | 5 | 5 | 0 | 0 | 0 | 0 | 5 | 5 | 0 | 0 | 0 | 0 |
| Male | 62 | None | None | 5 | 5 | 0 | 0 | 0 | 1 | 5 | 0 | 4 | 0 | 0 | 0 | 5 | 5 | 0 | 0 | 0 | 0 |
| Female | 56 | Father | Mother | 5 | 5 | 0 | 0 | 0 | 1 | 5 | 5 | 0 | 0 | 0 | 0 | 5 | 5 | 0 | 0 | 0 | 0 |

**Defining Ethnicity**

You might notice that there is no variable in the data which explicitly indicates the ethnicity of the subject. To serve the purpose of this survey – visualizing the vitality of languages in a community, ethnicity is defined solely by the linguistics competence of a subject’s guardians.  
To give a specific example, let subject A has a mother who **can speak**[2](https://liao961120.github.io/2019/02/17/visualize-language-loss.html#fn2) Mandarin and Taiwanese and a father who speaks Mandarin and Hakka, then subject A is categorized as a Mandarin, a Taiwanese, and a Hakka simultaneously.

The function filter\_ethnic() is used to filter out subjects with specified ‘ethnicity’. This function is useful for drawing age-sex pyramid for each of the six languages.

filter\_ethnic <- function(df, lang, lev = 3) {

sp\_lang <- vector("character", 3)

sp\_lang[1] <- paste0("dad\_", lang, "\_speak")

sp\_lang[2] <- paste0("mom\_", lang, "\_speak")

sp\_lang[3] <- paste0(lang, "\_speak")

df2 <- df %>%

filter(m\_guard\_identity != 'None' | f\_guard\_identity != 'None') %>%

filter(.data[[sp\_lang[1]]] >= lev | .data[[sp\_lang[2]]] >= lev) %>%

select(age, gender, sp\_lang)

return(df2)

}

**Assinging Age Group to Subjects**

Remember that the data structure needed for plotting age-sex pyramids requires **age group to be the basic unit**. data now consisits of single subjects, and we need information to group subjects together according to their ages. mutate\_age\_group() creates a new variable age\_group by the subject’s age.

mutate\_age\_group <- function(df, range = 5){

df$age\_group <- as.character(

cut(df$age, right = F, breaks = seq(10, 95, by = range))

)

return(df)

}

cut() takes a numeric vector as its first input, and codes the values of the vector into new values according to the interval they fall into. In mutate\_age\_group(), cut() codes the input vectors according to the intervals specified in the argument breaks.

Now we can use these functions to add more information to the data frame. The idea is to first create a separated data frame for each ethnicity[3](https://liao961120.github.io/2019/02/17/visualize-language-loss.html#fn3) (by filter\_ethnic()), then attact new variables to the data frame that indicate a subject’s ethnicity (ethn\_group) and age group (age\_group).

lang <- c('Mand', 'Tw', 'Hak', 'Ind', 'SEA', 'Eng')

lev <- c(3, 3, 3, 3, 3, 0)

ethn\_list\_df <- vector("list", length(lang))

for (i in seq\_along(lang)){

ethn\_list\_df[[i]] <- filter\_ethnic(data,

lang = lang[i],

lev = lev[i]) %>%

mutate(ethn\_group = lang[i]) %>%

mutate\_age\_group() %>%

select(age, gender, age\_group, ethn\_group,

paste0(lang[i], "\_speak")) %>%

rename(lang\_fluency = paste0(lang[i], "\_speak"))

}

Then we can recombine these data frames back to a single one, and this new data frame now has information about a subject’s ethinicity and age group he/she belongs to. (Note that the new data frame is **expended** since one subject can have several ethnicity, i.e. one subject can appear in different rows of the data frame with different ethnicity.)

bind\_rows(ethn\_list\_df) %>% head()

| **age** | **gender** | **age\_group** | **ethn\_group** | **lang\_fluency** |
| --- | --- | --- | --- | --- |
| 56 | Female | [55,60) | Mand | 4 |
| 57 | Female | [55,60) | Mand | 5 |
| 37 | Male | [35,40) | Mand | 5 |
| 44 | Female | [40,45) | Mand | 5 |
| 56 | Female | [55,60) | Mand | 5 |
| 53 | Male | [50,55) | Mand | 5 |

**Data for Plotting**

Finally, we are ready to group the subjects together according to his/her gender, ethnicity, and age\_group. After grouping, we can use dplyr::summarise() to calculate each group’s fluency of the language, which will be used as the variable on the horizontal axis of the age-sex pyramid.

pl\_data <- bind\_rows(ethn\_list\_df) %>%

group\_by(gender, ethn\_group, age\_group) %>%

summarise(mean(lang\_fluency)) %>%

rename(avg\_fluency = `mean(lang\_fluency)`)

head(pl\_data)

| **gender** | **ethn\_group** | **age\_group** | **avg\_fluency** |
| --- | --- | --- | --- |
| Female | Eng | [15,20) | 1.750000 |
| Female | Eng | [20,25) | 2.948718 |
| Female | Eng | [25,30) | 2.600000 |
| Female | Eng | [30,35) | 3.000000 |
| Female | Eng | [35,40) | 1.666667 |
| Female | Eng | [40,45) | 2.125000 |

**Plotting Function**

We are going to draw 6 age-sex pyramids, one for each languages. So instead of writing ggplot() six times, I wrote a plotting function:

library(ggplot2)

pl\_pyramid <- function(data, title = NULL) {

ggplot(data,

aes(x = age\_group,

y = ifelse(gender == 'Male', avg\_fluency, -avg\_fluency),

fill = gender)) +

geom\_bar(stat = "identity", width = 0.7) +

scale\_y\_continuous(limits = c(-5, 5),

breaks = seq(-5, 5, 1),

labels = abs(seq(-5, 5, 1))) +

coord\_flip() +

scale\_fill\_manual(values = c("#E41A1C", "#377EB8"),

breaks = c("Female", "Male")) +

labs(x = "Age", y = "Fluency", fill = "", title = title)

}

Now we can start plotting. Let’s try Mand (Mandarin) and Hak (Hakka) first. We can use dplyr::filter() to filter out people speaking these languages:

tweak <- theme\_bw() +

theme(axis.text = element\_text(size = 15),

legend.justification = "right",

legend.position = "bottom",

legend.box = "vertical")

pl\_data %>%

filter(ethn\_group == 'Mand') %>%

pl\_pyramid(title = 'Mandarin') + tweak

pl\_data %>%

filter(ethn\_group == 'Ind') %>%

pl\_pyramid(title = 'Formosan Languages') + tweak

Wait! It seems quite strange. The age-sex pyramid of ‘Formosan Languages’ doesn’t look like a pyramid at all!  
This is because there were very few subjects defined as ‘indigenous people’ in the survey. To make plots like this (with only one or two age groups) comparable to others (such as that in ‘Mandarin’), we need one more function to **insert missing age groups** to the data frame so that ggplot can draw empty bars for us.

First, we need to find out **all age groups in the data**:

age\_groups <- unique(pl\_data$age\_group)

age\_groups

[1] "[15,20)" "[20,25)" "[25,30)" "[30,35)" "[35,40)" "[40,45)" "[45,50)"

[8] "[50,55)" "[55,60)" "[60,65)" "[65,70)"

Then we can write a function fill\_empty\_age\_group(), which takes a data frame as its first argument and checks whether there are age groups missing in the data frame (using its second argument, age\_group\_all, as comparison). If the age group is missing, fill\_empty\_age\_group() appends a new row, which has age\_group set to the missing age group and avg\_fluency set to 0 (create an empty bar), to the input data frame.

fill\_empty\_age\_group <- function(df, age\_group\_all) {

for (i in seq\_along(age\_group\_all)) {

if (!(age\_group\_all[i] %in% df$age\_group)) {

df <- rbind(df, list(gender = "Female",

ethn\_group = "doesnt\_matter",

age\_group = age\_group\_all[i],

avg\_fluency = 0))

}

}

return(df)

}

Now we’re ready to explore language loss in Taiwan.multiplot() is used to put multiple plots together. The source code of multiplot() is copied directly from Winston Chang’s [Cookbook for R](http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)).

tweak2 <- theme\_bw() +

theme(legend.justification = "right",

legend.position = "bottom",

legend.box = "vertical")

py1 <- pl\_data %>%

filter(ethn\_group == 'Mand') %>%

fill\_empty\_age\_group(age\_groups) %>%

pl\_pyramid(title = 'Mandarin') + tweak2

py2 <- pl\_data %>%

filter(ethn\_group == 'Tw') %>%

fill\_empty\_age\_group(age\_groups) %>%

pl\_pyramid(title = 'Taiwanese') + tweak2

py3 <- pl\_data %>%

filter(ethn\_group == 'Hak') %>%

fill\_empty\_age\_group(age\_groups) %>%

pl\_pyramid(title = 'Hakka') + tweak2

py4 <- pl\_data %>%

filter(ethn\_group == 'Ind') %>%

fill\_empty\_age\_group(age\_groups) %>%

pl\_pyramid(title = 'Formosan Languages') + tweak2

py5 <- pl\_data %>%

filter(ethn\_group == 'SEA') %>%

fill\_empty\_age\_group(age\_groups) %>%

pl\_pyramid(title = 'Languages of South East Asia') + tweak2

py6 <- pl\_data %>%

filter(ethn\_group == 'Eng') %>%

fill\_empty\_age\_group(age\_groups) %>%

pl\_pyramid(title = 'English') + tweak2

multiplot(py1, py2, py3, py4, py5, py6, cols = 2)

**Language Loss in Taiwan**

As described in [Age-Sex Pyramid of Language](https://liao961120.github.io/2019/02/17/visualize-language-loss.html#age-sex-pyramid-of-language) above, we can learn about a language’s vitality in Taiwan from the age-sex pyramids drawn above.

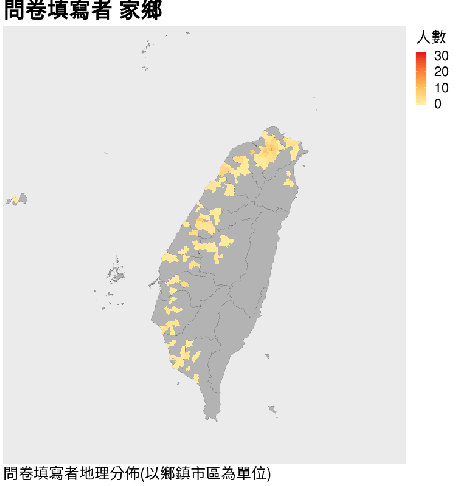
For Formosan and South East Asian languages, there are too few data, and we can’t learn much about these languages from the plots. For Mandarin, the shape of the pyramid is retangular, indicating the linguistic competence of Mandarin is stable across people of all ages. This is expected as Mandarin, in Taiwan, is the most prevalent language, and most people use it as the primary language in workplace and home.

Pyramids of Taiwanese and Hakka have inverted trianglular shapes, indicating lower linguistic competence among yonger people. Indeed, it’s not uncommmon to see the fathers and mothers talk to their children in Mandarin but talk to their parents in Taiwanese. Taiwanese is the second prevalent language in Taiwan, but it is shrinking, particularly among young people. Hakka ranks third in prevalence. It faces similar situation to Taiwanese but the situation is even worse, since the usage of Hakka is mostly restricted to Hakka people whereas usage of Taiwanese is not restricted to particular groups of people (many indigenous and Hakka peoples speak fluent Taiwanese).

English has a special status in Taiwan. It is not a native tongue to people born and raised in Taiwan, but many people know at least a little English. This is because formal education and the (awareness of) globalization lead parents to place importance on English education of their children. This also gives the age-sex pyramid of English its appearance – a triangular shape with wider bottom than top. English is the only language that is growing in Taiwan and, arguably, the language with the strongest vitality. [Table 1](https://liao961120.github.io/2019/02/17/visualize-language-loss.html#tab:tw-pyramids) summarizes the discussion above.

| Table 1: Situations of the four most prevalent languages in Taiwan. | | |
| --- | --- | --- |
| **Vitality of Language** | **Shape of Pyramid** | **Examples (Taiwan)** |
| Shrinking and Dying | Inverted Triangle | Taiwanese, Hakka |
| Growing | Triangle | English |
| Stable | Retangualr Bar | Mandarin |

1. The samples are not representative though, and it might only reflect the situation in Taipei (see the geographical distribution of the samples below). But I think there are still reasons to believe that other locations in Taiwan have similar phenomena.

[](https://liao961120.github.io/2019/02/17/visualize-language-loss.html#fnref1)

1. Defined by scoring 3 or above in a self-reported 6-point scale measuring the fluency of a language spoken by the subject’s parents.[](https://liao961120.github.io/2019/02/17/visualize-language-loss.html#fnref2)
2. For English, different from all other languages, the level used to determine ethnicity is 0, i.e. there is no filtering occuring, and all subjects are used. This is because, in Taiwan (and many other non-English speaking communities as well), English is not related to ethnicity and is strongly related to formal education and job requirements.[](https://liao961120.github.io/2019/02/17/visualize-language-loss.html#fnref3)